DISSERTATION

MASTERY QUIZZING: ASSESSING A NOVEL TESTING TECHNIQUE IN THE CLASSROOM AND THE LABORATORY

Submitted by

Lauren Elizabeth Bates

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2019

Doctoral Committee:

Advisor: Edward DeLosh

Matthew Rhodes
Daniel Graham
James Folkestad

ABSTRACT

MASTERY QUIZZING: ASSESSING A NOVEL TESTING TECHNIQUE IN THE CLASSROOM AND THE LABORATORY

Research promotes the use of frequent quizzing, as well as the use of feedback to promote long-lasting learning. In this dissertation, I propose a method for promoting long-lasting learning using mastery quizzing. Participants read an expository text and then answered questions about that text. Some participants were required to take quizzes until they achieve a perfect score, which I refer to as mastery quizzing, whereas other participants were forced to take quizzes a certain number of times. I explored how mastery quizzing can contribute to students' classroom learning and whether this method is more effective than traditional quizzing. In Experiment 1 I first looked at whether the benefits of mastery quizzing may emerge due to the benefits associated with frequent testing and feedback. Next, In Experiment 2 I explored the role that feedback may play in the mastery model, exploring students' use of feedback and how that may impact final test scores. Experiment 3 explored whether attending to and processing feedback led to increased performance on a final test. My results supported an overall benefit of mastery quizzing relative to standard quizzing techniques, even when controlling for number of quiz attempts, the presence of feedback, and conditions meant to simulate a need to use the feedback to improve performance. These results imply that the mastery technique may be a more effective method to improve student learning than standard quizzing techniques.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

A large body of research on human learning and memory has consistently revealed that taking a practice test on information relative to simply restudying it leads to better long-term retention of that information. This phenomenon is known as the testing effect (Roediger & Karpicke, 2006b). One recent meta-analysis showed that the probability of remembering information on a final memory test was 2.5 times greater after practice testing than after an equivalent amount of time spent restudying (mean effect size of 0.55; Rowland, 2014). Thus, quizzes and tests are not only useful for assessing learning, they can also be quite effective for enhancing learning as well. What this suggests is that learning does not occur by simply taking information in but that the process of retrieving a memory may be a more significant contributor to long-lasting learning than other methods of reviewing that information.

Before exploring different theories of the testing effect, it is important to outline how testing effects are studied. In a typical testing effect paradigm, participants are first presented with information in what is deemed the learning phase. This information may comprise lists of words, maps, prose passages, or names to provide a few examples. Next, participants enter an intervening phase in which they either restudy the material, engage in practice testing, or restudy/test on different material. After a predetermined retention interval, participants are given a final test on their memory for the items they previously encountered in the experiment. A testing effect is evident if a testing condition yields better final test performance than a restudy condition, as this suggests a memory benefit for testing relative to restudying.

Theories of the Testing Effect

Investigating the efficacy of the testing effect requires an understanding of the mechanisms that underlie retrieval. The goal when retrieving something from memory is to use that information to shape current thought and behavior (Kuhl & Wagner, 2009). Research on the testing effect suggests that the process of retrieval itself strengthens memory. This strengthening is aptly demonstrated by studies that utilize multiple retrieval attempts. For instance, Roediger and Karpicke (2006b, Experiment 2) compared repeated study to different testing manipulations (i.e., one in which participants had one practice test and another in which participants had three practice tests). All three experimental conditions were matched for time and number of exposures to the to-be-learned material. Participants given three practice tests had the greatest long-term retention on a final memory test administered one week later, followed by those with one practice test and then those in the repeated study condition. These data indicate that the benefits associated with retrieval practice may not be solely due to the increased amount of time spent with the material. Where the following theories differ is in how they explain the underlying mechanism(s) of retrieval that strengthens a memory.

Transfer Appropriate Processing (TAP)

The transfer appropriate processing (TAP) account of the testing effect suggests that the testing effect may stem from the high degree of similarity between the practice test phase and the final test phase, implying that the benefit of retrieval practice is related to its structural similarity to the final test. That is, the TAP account assumes that the practice test condition shares more features with the final test than the restudy condition. Because participants engaging in retrieval practice are practicing the task they will ultimately do on the final test, the TAP account states that this would explain their higher performance on the final test. That is, because the practice

test is highly similar to the final test, this aids in retention (Roediger & Karpicke, 2006a). Therefore, similarity across conditions is emphasized as a key feature rather than retrieval effort. TAP predicts that testing effects should be larger if there is more similarity between the initial test and the final test.

Some empirical evidence has emerged to support TAP (Roediger & Karpicke, 2006b), however there is also evidence to suggest that it does not hold up when empirically tested (Carpenter & DeLosh, 2006; Carpenter, 2009). Evidence in support of a TAP account demonstrates that learners perform best on a final test when the final test is highly similar to their initial practice test. For example, McDaniel and Fisher (1991) had participants take an initial test on factual questions and then found that, on a later test, their performance was better when they were tested on the same questions relative to reworded questions on the same topic. Thus, the high degree of overlap between what participants experienced during the initial test and what they experienced during the final test benefitted their long-term retention.

Arguments against TAP suggest that conditions exist under which successful long-term retention depends on other factors beyond the degree of overlap between initial test and final test format. For example, some experiments have shown that, regardless of final test format, free recall serves as the best initial test as it often leads to the best final memory performance. Some research has even shown that more closely matching the initial and final tests does not necessarily lead to a testing effect (Carpenter & DeLosh, 2006). Importantly though, a meta-analysis revealed that a match between initial and final test did not yield a reliable change in magnitude of the testing effect (Rowland, 2014). Additionally, TAP simply offers a re-description of the testing effect, without identifying an underlying mechanism. Therefore, TAP alone cannot serve as a viable explanation of the testing effect.

Elaborative Retrieval/Retrieval Effort

Darley and Murdock (1971) originally suggested that recall aids the accessibility of memory, theorizing that having prior exposure to recall makes it easier to access the recalled information on a later test. Modern theoretical accounts of the testing effect also suggest that the act of testing changes the accessibility of a memory (Carpenter, 2009; Carpenter, 2011; Karpicke, Lehman, & Aue, 2014). For instance, the elaborative retrieval hypothesis (ERH) suggests that testing is beneficial because, during the retrieval process, the learner activates information related to the target response (i.e., semantic, or meaning-based, aspects of items). This increases the chances that activation of any of this information will later facilitate retrieval of the target, creating a "web" of spreading activation that would likely not appear in a restudy condition (Carpenter, 2009; 2011; Carpenter & DeLosh, 2006). For example, Carpenter (2009) had participants study cue-target word pairs, with cues manipulated to either be strongly (e.g., Basket – Bread) or weakly (e.g., Toast – Bread) associated with their targets. On the initial test, strong cues were better recalled than weak cues and participants took significantly longer to recall weak cues. Weak cues are thought to be more difficult and therefore require a more extensive search during retrieval. However, on a final test after a five-minute delay, weak cues led to better retention than strong cues. This suggests that retrieval effort may be directly related to endurance of the memory.[1]

[1] It also bears some relationship with work on retrieval fluency, the ease in which one can retrieve information (Soderstrom & Bjork, 2015). The work on retrieval fluency demonstrates that just because a response is more easily recalled initially does not mean it is better remembered (Bjork, 1999; Soderstrom & Bjork, 2015). That is, what may be perceived as more fluent or available in memory can lead one to falsely assume that such information will be more likely to be remembered in the future. Instead, difficulty retrieving something can make that information more likely to be remembered in the future, assuming an initial retrieval attempt is successful. Therefore, having a challenging retrieval scenario is better for long-term retention of information, even though it may not make the learner feel as successful if they find they are struggling to come up with information.

Further explanations of the ERH have considered the role of lag (i.e., the time interval in-between practice test trials) in explorations of the testing effect. For instance, several studies have demonstrated that increasing the lag before a practice test increases the difficulty of retrieval, thereby strengthening the memory. Rawson, Vaughn, and Carpenter (2015) explored this by manipulating how many trials were in between items that participants would ultimately be tested on. In the long-lag groups, each block of trials included all 36 items. Thus, trials for a given item were separated by 35 other items. In the short-lag groups, items were randomly assigned to four sets of nine pairs, and all six blocks of trials for one set were presented before proceeding to trials for the next set. Thus, the trials for a given item were separated by eight other items. They found that there was an interaction between lag and retrieval practice, such that participants benefitted more from retrieval practice when the degree of lag was longer. The long lag group had a greater benefit of practice testing than the short lag group. Therefore, retrieval practice may have its memory improvement effects enhanced by increasing the amount of time in-between to-be-learned information.

Pyc and Rawson (2009) sought to test this retrieval effort hypothesis by manipulating the interstimulus interval (ISI) (i.e., the amount of time that passes between the presentation of items) and criterion level (i.e., the number of times items were required to be correctly recalled before dropping from practice). They predicted that a longer ISI should lead to better long-term retention since this would be more difficult, supporting the retrieval effort hypothesis. In addition, they assumed that as the number of times items that are correctly retrieved increases, the incremental benefit to final test performance would decrease. Their findings supported the retrieval effort hypothesis, showing that successful but difficult retrievals were most beneficial for long-term retention. Additionally, their criterion level manipulation supported the utility of

repeated retrieval, albeit they found diminishing returns as the number of trials to criterion increased.

A recent meta-analysis showed that many of the predictions of retrieval effort theories of the testing effect were upheld (Rowland, 2014). The type of initial test influenced the magnitude of the effect size, such that more difficult tests (e.g., free recall relative to recognition) led to greater effect sizes. Free recall tests are generally more difficult than other types of test because they require learners to produce an answer with very few cues to help trigger their memory. For example, asking a question such as, "Which school of thought did psychologist Wilhelm Wundt develop?" is more difficult than asking the same question in a multiple-choice format whereby learners can then choose from a list of options when deciding on the correct answer. Rowland's meta-analysis further showed that more difficult initial retrieval attempts led to a greater testing effect. Thus, with some exceptions (e.g., there was no relationship between number of retrieval attempts and the magnitude of the testing effect), retrieval effort largely gathered support from this meta-analysis.

As the name suggests, the retrieval effort hypothesis suggests that effort is a key contributor to the magnitude of the testing effect. That is, the more difficult the initial retrieval opportunity, the better the memory performance would be on a final test (Carpenter & DeLosh, 2006). This has been shown empirically within the testing effect research, but also has wider implications for work on desirable difficulties (Bjork, 1999). Making the learning environment more difficult has been shown in many cases to enhance memory. This can help explain the testing effect's memory benefit over restudying, but it can also explain other powerful memory phenomena such as the spacing effect (Soderstrom & Bjork, 2015). Even studies on transfer of learning have championed effortful processing (Salomon & Perkins, 1989; Barnett & Ceci,

2002). In contrast, low effort (such as simply reviewing material) may foster an illusion of learning (Soderstrom & Bjork, 2015).

Although the role of effort in the testing effect is one important consideration, it is not without its own caveats. An important exception has to do with individual differences and task differences. Some research has shown that more effortful learning techniques may be less efficient than less effortful techniques, or even harmful in some situations, such as when individuals have different levels of knowledge. For instance, Carpenter et al. (2016) demonstrated that while high knowledge learners benefit more from effortful strategies, low knowledge learners may not benefit. Additionally, research in the motor learning literature suggests that complex skills may require less effortful training due to the large amount of demand placed on the learner (Wulf & Shea, 2002). While these findings do not discredit the importance of effort in retrieval practice, they suggest some caution in assuming testing effects reflect retrieval effort.

Elaboration is the extent to which learning is enriched by integrating information and forming associations. Studies on elaboration in many areas of cognitive psychology often show that it is implemented when engaging in a difficult task (Hughes & Whittlesea, 2003; Wulf & Shea, 2002; Mulligan & Lozito, 2004). Applied to testing, the ERH, holds that, when engaging in retrieval, mediators are activated to aid in memory. These mediators could be words, for example, that provide a link between the cue and the target (e.g., studying Basket – Bread, may activate related words, such as Basket – Eggs – Flour – Bread). The ERH presumes that when learners engage in retrieval practice, they activate other information related to the information they are trying to retrieve. Therefore, the more difficult it is to retrieve that information initially, the more connections made with items that are activated to remember the information in

question. For instance, items that are weakly associated, such as Basket – Bread, may involve more mediators than strongly associated pairs such as Toast – Bread. If mediators are important for long-term retention, then weakly associated cues should demonstrate better memory over time than strongly associated cues (Carpenter, 2009, 2011).

Thus, retrieval difficulty could foster the formation of mediators since it would require the learner to make more associations between items for weakly associated pairs. Pyc and Rawson (2010) further clarified this by stating that mediator retrieval (i.e., whether the mediator can be recalled when given a cue) and mediator decoding (i.e., whether the mediator correctly elicits the target item from memory) are necessary for effective mediators. They found that mediator retrieval led to better final test performance on a memory task when participants engaged in retrieval practice, relative to restudy practice. This supports the mediator effectiveness hypothesis and suggests that the testing effect may be driven by generating mediators during encoding. Successful retrieval of these mediators may then further strengthen memory for the information in question.

Pyc and Rawson (2012) later extended this work by exploring the mediator shift hypothesis, which proposes that failed retrieval attempts during practice testing allow the learner to evaluate the effectiveness of their mediators, thereby allowing them to shift to more effective mediators if necessary. Participants engaged in a keyword strategy to learn foreign language word pairs. This strategy required participants to generate an English word when learning Swahili-English word pairs that was like the Swahili word either semantically or phonetically (e.g., when learning the pair wingu-cloud, one might generate the word “wing” to remind them of birds that have wings and fly in the clouds). Across three experiments, the authors successfully demonstrated that when engaging in retrieval practice relative to restudy,

participants in their study were more likely to shift their keywords, especially after a failed retrieval attempt, to ultimately lead to a higher rate of retrieval success on a final memory test. These findings support the proposed mediator shift hypothesis, further contributing to the theme that the underlying mechanisms of the testing effect may reflect both the associative strength between learned items and mediators and the difficulty experienced during learning. For example, experiencing retrieval failure on a practice trial can prime the learner for more robust long-term retention since it requires greater effort, and therefore enhances the connection between cues and the to-be-learned information.

Carpenter and Yeung (2017) further explored the role that mediator strength plays in learning from retrieval by attempting to experimentally test its potential as an underlying mechanism of the testing effect. Specifically, they manipulated mediator strength to see if words with higher mediator strength benefitted more from retrieval practice. Mediator strength refers to the strength of association between a cue and a mediator. For example, according to norms established by Nelson, McEvoy, and Schrieber (2004), the pair Chalk – Board has a forward associative strength of 69%, meaning that when given the word Chalk, the word Board is produced 69% of the time. This would be an example of a strong mediator. To contrast, the word Soup produces the associate Chicken only 10% of the time, which would serve as an example of a weak mediator. Carpenter and Yeung (2017) demonstrated that during long lags, mediators were more likely to be activated and therefore more likely to lead to retrieval success in a testing effect paradigm. In addition, mediator strength was related to the testing effect such that those in the retrieval practice condition demonstrated a more robust testing effect than those in the restudy condition. When mediators were activated during retrieval, this aided later retention of the information. Mediator strength did not affect participants who only restudied information.

Despite the criticisms surrounding the mediator effectiveness hypothesis (Carpenter, 2011), it cannot be denied that elaboration seems to play a role in the testing effect. Even if there are other underlying mechanisms that have yet to be discovered, there is empirical support behind the use of elaboration. This largely comes from the retrieval effort hypotheses that mention the spreading of activation model (Carpenter, 2009; Carpenter, 2011). That is, activating information by way of retrieval subsequently activates related information that can later serve as memory cues. Therefore, the harder it is to retrieve that information initially, the more connections that are made with items that are activated to remember the information.

Some research has revealed that, in addition to direct benefits of retrieval on the retrieved information, the testing effect may contribute to enhanced memory for information that is not itself retrieved. Such retrieval-induced facilitation (RIFA) suggests that successfully retrieving information can also benefit information that is not retrieved but is somehow related to the retrieved information (Chan, 2009; Chan, McDermott, & Roediger, 2006; Rowland & DeLosh, 2014). This effect has emerged when considering non-retrieved information that is semantically related to the retrieved information (Chan et al., 2006; Chan, 2009) as well as information that is episodically related to the retrieved information (Rowland & DeLosh, 2014). Thus, the act of retrieval goes beyond simply aiding the information that was retrieved, as these studies also show, but goes on to enhance memory for related information that was not subjected to the test itself. Both veins of research suggest that the act of retrieval is beneficial to memory, though in different ways.

Episodic Context Accounts of Testing

Any comprehensive explanations of the testing effect will likely incorporate multiple processes to account for the benefit of retrieval on subsequent memory. The studies on retrieval

effort, specifically those that consider the role that elaboration plays in retrieval practice (Carpenter, 2009; Carpenter, 2011), suggest that the underlying mechanisms of the testing effect must relate to elaboration. However, findings from the initial study on the mediator effectiveness hypothesis (Carpenter, 2011) have been criticized as being correlational (Karpicke et al., 2014). In addition, although the ERH focuses almost exclusively on the importance semantic cues, more recent research considers whether episodic cues may better explain the mechanisms underlying the testing effect (Karpicke et al., 2014). According to this episodic context account, retrieval practice encourages the learner to reinstate the original learning context, leading to better memory on a final test (Whiffen & Karpicke, 2017). This is an important distinction to make, because previous theories are largely based on the level of connection between cue and target items in terms of meaning. The episodic context account assumes that retrieval permits the learner to think back to the original context in which they first encoded the information. This account does not necessarily discount the role that meaning plays in retrieval, but instead focuses on how another mechanism (i.e., encoding context) may be a significant contributor to retrieval success.

The episodic context account also explains effects such as the increased memory performance when engaging in spaced over massed retrieval practice (Roediger & Karpicke, 2011). In a test of the episodic context account, Whiffen and Karpicke (2017) had participants study words, then either restudy them or answer questions regarding whether the word they were being shown was on one of two earlier lists. This required participants to think back to the conditions under which they had originally learned the words. Final memory performance was better when participants had completed the list discrimination task, in which they had to think back to which list each item belonged to, compared to performance in the restudy condition.

In a subsequent experiment, Whiffen and Karpicke (2017) added a new condition in which participants made semantic judgments (i.e., pleasantness ratings) about the list items. Final memory performance was not significantly different when comparing the original list discrimination task to the pleasantness rating task, but both were superior to restudy. Additionally, the list discrimination task led to superior temporal organization of the materials. In a third experiment, participants studied categorized lists of words rather than the unrelated lists of words that had been used in the two previous experiments. This experiment also included a category judgment task in which participants had to identify the category that a word belonged to (e.g., if given the word banana, participants would judge between the categories of animals and fruits to determine which the word belonged to). Final memory performance was better for the list discrimination condition compared to restudy, but it was not significantly greater than the category judgment condition. When comparing output, those in the list discrimination condition relied on a temporal output strategy (i.e., recalling words based on list order), whereas those in the category judgment condition relied on a semantic output strategy (i.e., recalling words based on category membership). This dissociation reveals support for the episodic context account, because it shows that when one reinstates the temporal context in order to remember information, it is temporal cues that guide retrieval.

Discussion of Perspectives

In their meta-analysis and theoretical review of the literature, Chan, Meissner, and Davis (2018) found that two clusters of theories best explained the benefits of retrieval practice: theories that supported retrieval practice as a way of reducing proactive interference (PI) (i.e., old information interfering with new to-be-learned information), and theories that emphasized the importance of integration between items that are learned, specifically pointing out semantic

relatedness as a positive example. These two clusters of theories seemed to lead to more consistent testing effects relative to theories that suggested learners adjust their encoding strategies based on what they experience during retrieval practice, and theories suggesting that context at encoding and retrieval served as important cues for learning. That is not to say that the latter two theories do not have empirical support, but rather that their support does not appear to be as strong as support for theories that focus on integration and theories that focus on limiting PI.

In the context of this paper, this would suggest that ideas like the mediator effectiveness hypothesis and episodic context account contribute a great deal to understanding of the testing effect, and theories such as TAP do not account for as much of the effect. However, it is limiting to propose that there is one theory that can sufficiently explain the testing effect in its entirety. Rather, these differing theories can be thought of as exploring different aspects of a complex phenomenon. Also, several theories mentioned here differ in their underlying logic, making it hard to strictly compare them without considering potential contradictions between them. Thankfully, if one of the main factors contributing to the success of retrieval practice is semantic integration of materials, this lends itself well to the context of education, which is a field that is highly relevant to discuss when considering the efficacy of the testing effect.

Classroom Studies on the Testing Effect

Classroom research on the testing effect offers strong practical implications for the use of practice tests in a class setting (Roediger & Karpicke, 2006a; Roediger & Karpicke 2006b; Carpenter et al., 2016; Roediger, Agarwal, McDaniel, & McDermott, 2011; Carpenter, Pashler, & Cepeda, 2009; Goossens, Camp, Verkoeijen, & Tabbers, 2014; Jaeger, Eisenkraemer, & Stein, 2015; Karpicke, Blunt, Smith, & Karpicke, 2014; McDaniel, Anderson, Derbish, & Morrisette,

2007; McDaniel, Wildman, & Anderson, 2012). Carpenter and colleagues (2009) demonstrated that retrieval practice could improve students' memory for U.S. history facts and even offered support for the use of spacing in classroom settings. Roediger and colleagues (2011) found across three experiments that introducing low-stakes quizzing into a middle school curriculum not only improved students' chapter exam and final exam scores, but it also reduced some students' test anxiety. The benefit that testing has on long-term retention of material is something that has been robustly demonstrated in laboratory experiments as well, but to see it in a classroom setting serves as an important piece of evidence for the educational merits of this work. Furthermore, the effect that retrieval practice has on test anxiety could be an indirect effect of testing (i.e., a byproduct, or supplemental effects beyond the main effect that it has on long-term retention), to contrast with the direct effect that retrieval practice has on memory for the tested items (Roediger, Putnam, & Smith, 2011; Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014). The benefits of testing in the classroom appear to have an impact on more than just students' memory for the information in question.

Are students aware of the benefits of testing? Hartwig and Dunlosky (2012; see also Kornell & Bjork, 2007)) surveyed over 300 college students by assessing their study strategies and GPA, with the goal of learning more about which strategies are associated with higher achievement. Some of the most commonly used strategies were ones that are considered "low utility" (i.e., they are not as effective as engaging in active learning), and included things such as rereading notes, highlighting relevant text, and cramming the night before the test. Additionally, students reported frequent use of practice testing. Practice testing does serve as an active learning strategy, as it requires learners to attempt to retrieve information from memory. However, most students utilized testing as a metacognitive tool to assess their performance, rather than a

learning tool to improve their retention (Hartwig & Dunlosky, 2012). This implies that the learning benefits of retrieval practice may not be widely appreciated.

Morehead, Rhodes, and DeLozier (2016) sought to investigate both college students' and instructors' knowledge of effective study strategies. They found that instructors were only slightly better than students when it came to endorsing empirically supported study strategies, and that students frequently reported using ineffective study strategies (e.g., rereading material). In addition, students frequently chose what to study by looking at what was due next (see also Kornell & Bjork, 2007). Morehead and colleagues (2016) also pointed out a few interesting inconsistencies between what students and instructors in their study reported. For example, while most instructors mentioned that they reviewed study strategies in class, only 36% of students surveyed said they utilized study habits that they had learned about from an instructor. When considering the fact that instructors in their study demonstrated a slightly firmer grasp on evidence-based practices for learning than students, we should still consider instructors as a valuable resource for disseminating information about effective study skills.

Based on this body of literature, there are compelling reasons to have instructors test students more frequently in courses, not just for assessment purposes, but for enhancing learning. This could come in the form of weekly quizzes, as these have been shown to demonstrate a benefit of learning, and even increase students' enjoyment and satisfaction with the class (Leeming, 2002). For actions that students could take, this could manifest as use of sample test questions, flashcards, or taking turns testing each other with fellow classmates. There are many ways one can incorporate these strategies into the classroom. I propose a method that incorporates both these benefits involved with frequent testing as well as those associated with another phenomenon: feedback.

Investigating the Role of Feedback

The literature on feedback demonstrates that the benefits associated with retrieval practice are enhanced when learners are provided with feedback (Roediger & Butler, 2011; Rowland, 2014). If learners are given the correct answer after a retrieval attempt, it allows them to correct their errors on future retrieval attempts (Pashler, Cepeda, Wixted, & Rohrer, 2005; Kulik & Kulik, 1988). Feedback is especially critical when learners are taking multiple choice (i.e., recognition) tests, since incorrect information is being provided to them along with the correct information (Butler & Roediger, 2008).

Dunlosky and Rawson (2015) sought to experimentally investigate whether students utilize testing as a learning tool, and if so, whether they used the feedback from their practice testing to help them on later retrieval attempts. More relevant for the present discussion, they also sought to understand whether students would continue to test themselves, with feedback, until they had the ability to correctly recall the information they had learned one or more times. Dunlosky and Rawson (2015) had participants in the self-regulated learning (SRL) group to choose how they wanted to study the material they were learning in the experiment. Participants could either elect to restudy a concept, take a practice test, or make a judgment about how well they would remember the information later. They also had the ability to select when they were done practicing for the eventual final test. The findings from their study demonstrated that, when engaging in retrieval practice, students typically used feedback from the practice test trials as a learning aid. However, this result was more likely when students were incorrect on a given question. When compared to other groups in which participants had to correctly recall the information one or three times, those in the SRL group tended to have lower overall performance when considering scores on the practice tests. Final test scores revealed that those in the SRL

group tended to perform worse than those in the other groups in which they were forced to continue to practice test until they had achieved a certain criterion (i.e., one or three correct retrievals). Thus, students may not actively participate in best practices for learning and may instead benefit more when placed in a scenario in which they are required to complete a certain number of practice tests.

Dunlosky and Rawson (2015) also revealed that students tended to seek out feedback after a relatively short delay. This implies that students are aware of the benefits of feedback as a learning tool and wanted to utilize it relatively quickly after taking a practice test. Additionally, they found that students who learned to criterion (i.e., continued to study until they had perfect performance on the given memory task) had greater retention on their final test, but achieving a perfect score was not required; therefore, not every student continued to test themselves until they had learned to criterion. They tentatively concluded that learning to criterion led to better long-term retention, but with students in control of the decision to continue testing themselves, individual differences (e.g., background knowledge, differential forgetting) may have contributed to this finding (see also Hays, Kornell, & Bjork, 2010).

This study presents several interesting ideas; students may intuitively know that feedback is a useful tool when learning but may not fully utilize it if they are not motivated to learn to criterion. Studies on students' metacognitive awareness of their own learning of complex materials reveal that students are not good at judging their level of understanding of the information (Dunlosky & Lipko, 2007), though some work suggests that a delay may improve students' metacognitive judgments (Pyc, Rawson, & Aschenbrenner, 2014; Rhodes & Tauber, 2011; Anderson & Thiede, 2008). Therefore, it may be more helpful to require learning to

criterion rather than let students make a choice. Perhaps this forces students to attend to feedback, thereby increasing their chances of utilizing its benefits.

The Present Study

I suggest a potential solution for incorporating frequent testing with feedback into courses, one that follows from empirical research on practice testing while simultaneously addressing barriers and concerns related to some of the practical aspects of frequent quizzing (e.g., use of class time, the grading burden). A viable method of testing may be one that requires students to complete quizzes outside of class using an online learning management system (e.g., Blackboard, Canvas). A unique characteristic of this quizzing procedure is that it employs a method that I will refer to as a mastery model of testing, as the method mandates that students must get all the answer correct. That is, instructors identify a set of key concepts that they expect all students to learn and students must demonstrate mastery of all concepts to get credit for this type of quiz. Students may take the quiz as many times as needed up until a deadline, but they must get all questions correct on one of their quiz attempts to receive credit for the quiz. That is, credit for the quizzes is assigned in an all-or-none fashion dependent on whether students demonstrate full mastery of the concepts. As such, students engage in repeated testing, receive feedback after each attempt, and are the need for mastery motivates them to attend to that feedback.

In a typical implementation of this mastery model of quizzing, I have given weekly quizzes comprised of 10 multiple-choice questions. For each week in which a quiz is given, the quiz is deployed online, and students are tested on ten key concepts from that week's material. Students must demonstrate mastery of all ten concepts by a deadline by taking and re-taking the quiz until they get all the questions correct on one of their quiz attempts. To ensure that students

are not simply memorizing answers, the questions change from one instantiation of the quiz to the next. For each concept, I have written two or three questions that pertain to that concept. One question on classical conditioning might ask, for example, “In Pavlov’s classic experiment, what was the conditioned stimulus?” A different question on the same concept might state, “Let’s say you’ve long had the habit of eating junk food while watching TV. Now you find yourself suddenly feeling hungry when you turn on the TV. What has the TV become, using classical conditioning terminology?” The online management system randomly selects one question from each concept set when an instantiation of a quiz is delivered, ensuring that the specific mixture of questions differs from one quiz attempt to the next. The set of questions is also given in random order (in the case of Blackboard) and the response options are scrambled as well (both Blackboard and Canvas).

Immediately following each quiz attempt, the online management system automatically delivers detailed feedback to students. Students are not only shown the original questions and correct answers, but they also get a detailed explanation (input into the system by the instructor) as to which answer is correct and why. Thus, students can correct misconceptions and identify areas that need further study prior to taking the quiz again since they have been provided with feedback.

Although potentially promising, I am unaware of any experiments (a) testing the efficacy of this approach and (b) comparing mastery quizzing to other approaches to quizzing. Accordingly, I ran several experiments to test the mastery model and explore different aspects of this technique. In this first experiment, I compared four conditions: having one attempt at a quiz, having two attempts at a quiz, having three attempts at a quiz, and using the mastery model (i.e., as many attempts as are required to earn a perfect score). Past research has demonstrated that

numerous retrieval practice attempts may lead to better final test performance (Roediger & Karpicke, 2006b). For this reason, I hypothesized that final test performance would increase as number of quiz attempts increased.

Additionally, there is considerable research demonstrating the benefit that feedback has on long-term retention (Roediger & Butler, 2011; Pashler et al., 2005; Butler & Roediger, 2008; see Kluger & DeNisi, 1996, for a review of feedback). Therefore, I also investigated the hypothesis that the mastery model would be a superior quizzing method because it implements frequent quizzing and feedback. I predicted that while there would be a small increase in final test performance between the first three conditions due to the influence of repeated retrieval practice, there would be a sharp increase in performance when comparing these standard quizzing conditions to the mastery condition since the mastery condition includes feedback. Stemming from this idea, I investigated the critical role of feedback in the mastery model in Experiments 2 and 3. One explanation regarding the efficacy of the mastery model is that learners are required to attend to feedback to improve their performance on the practice tests. Experiment 2 included three conditions: one attempt at a quiz, two attempts at a quiz, and the mastery model. Unlike in Experiment 1, all three conditions included feedback to more closely examine whether a benefit of mastery quizzing would disappear when controlling for the presence of feedback. I hypothesized that the mastery quizzing method would be superior to the other two conditions since it included the potential for more frequent retrieval, as well as an implied requirement for participants to attend to and process the feedback, as they were required to achieve a perfect to move on to the next phase of the experiment. I also hypothesized that including feedback in the other two conditions would reduce the benefit of mastery quizzing that I anticipated in Experiment 1, but that mastery would still be the superior technique.

In Experiment 3, I approached this question differently by comparing three conditions: one attempt at a quiz with detailed feedback given, one attempt at a quiz with detailed feedback given followed by a rating question (i.e., "On a scale of 1 - 5 with 1 being 'very poor' and 5 being 'very good' rate your knowledge on the concept you just learned about"), and the mastery model. This was designed to explore if the benefit of mastery quizzing stemmed from learners' need to look at feedback to correct past mistakes. Experiment 3 may reveal whether the benefit of mastery quizzing reflects the need to fully attend to and process the feedback given, since the nature of the rating task implies that the learner must assess their own knowledge before continuing. Therefore, I hypothesized that mastery would still demonstrate an overall benefit as a technique that implements frequent testing. However, when controlling for number of attempts, the benefit of mastery would disappear when compared to the condition that included a rating task, as this was meant to simulate learners' need to effectively attend to and process their feedback before moving on to the next phase of the experiment.

CHAPTER 2: EXPERIMENT 1

Experiment 1 examined whether the benefit of mastery quizzing is largely due to the benefit of multiple practice tests and feedback, given previous research showing that multiple retrieval practice attempts and feedback lead to superior long-term retention. This was accomplished by comparing the mastery quizzing method to conditions with one quiz attempt, two quiz attempts, or three quiz attempts (with no feedback). As such, the experiment allows a direct comparison of mastery quizzing to the standard quizzing procedure that is commonly used in classrooms, but also examines whether multiple quiz attempts using a standard quizzing method (with no feedback) produces performance comparable to mastery quizzing. If these standard quizzing conditions lead to comparable performance when controlling for the number of attempts in the mastery condition, we could reasonably conclude that any benefit the mastery model may offer would be more related to the frequency of testing rather than the inclusion of feedback.

Method

Participants

After conducting a power analysis using G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007), the required sample size for detecting a small effect (i.e., d = .25) with .8 power was 180 participants. Participants were recruited using Amazon Mechanical Turk and were compensated for their time at the rate of $1.00 for every 10 minutes. The survey time was set at 40 minutes in length, meaning compensation for each participant was $4.00. Due to an error in

programming, I had to re-run the mastery quizzing condition and decided to increase the time made available to participants to 50 minutes, thereby increasing the incentive to $5.00.

Of the 180 participants, 20 did not complete the experiment. The initial random assignment of participants to conditions led to there being 39 participants in the one quiz condition, 40 participants in the two-quiz condition, and 38 participants in the three-quiz condition. Due to the programming error, 43 participants ended up in the mastery condition, leading to a total of 160 participants. Thankfully, this distribution does not reflect the pattern of participant attrition that can be found in experiments that use online samples, as attrition did not differ greatly among the four conditions (Zhou & Fishbach, 2016).

Design

This experiment included four between-subjects conditions: one quiz attempt, two quiz attempts, three quiz attempts, or the mastery model (i.e., as many quiz attempts as are necessary to achieve a perfect score). All participants in the first three conditions were permitted 40 minutes to complete this experiment, and participants in the mastery condition were permitted 50 minutes to complete this experiment. After piloting the experiment with several psychology undergraduate students and non-psychology faculty members, I chose this time to allow for non-experts to complete the task without feeling rushed. The measure of interest was participants' scores on the final test. Each participant received a final test that was composed of ten questions. Therefore, participants' final test performance reflects the number of questions they answered correctly out of 10.

Due to the nature of this experiment, participants in the mastery model were required to complete the experiment within the time constraint of 50 minutes. Participants were also limited to nine mastery quiz attempts. That is, participants were automatically sent to the next phase of

the experiment if they needed nine or more quizzing attempts to achieve mastery. This was set in place to account for participants potentially running out of time within the 50-minute window. These factors may limit the generalizability of the findings as typical students may choose to spread their quizzing attempts out over several days if given a mastery quiz that must be completed during a week. However, note that finding an effect of mastery in this experiment would still be relevant because it would limit the amount of spacing that participants receive. Spacing out the quizzes should make participants' performance higher (Roediger & Karpicke, 2011); thus, asking participants to engage in mastery quizzing in a limited time may lead to a smaller memory effect than one might see in a classroom setting. However, if such an effect exists, it implies that there is a possibility of a stronger effect in a more realistic setting.

Materials

Quiz questions were based on concepts from an introductory psychology textbook (Myers & DeWall, 2016). Quizzes were administered using Qualtrics. Each quiz contained a total of ten questions on ten separate psychology concepts (i.e., the brain, learning, memory, personality, social psychology, child development, research methods, the nervous system, sensation and perception, abnormal psychology). See Appendix A for the complete set of materials from the learning phase.

I created four questions for each concept. The first three questions were used for the quizzes (e.g., the first question on the cerebral cortex appeared in the first quiz, the second question on the cortex appeared in the second quiz, if applicable, and so on). The fourth question was used on the final test at the end of the experiment and was always an applied question (i.e., one that required participants to go beyond simply defining the term they were learning). Each question was given as a multiple-choice question with four options, with the order of the options

randomized on each presentation. All quiz and final test questions from this set of experiments can be viewed in Appendix B.

In this experiment, participants only received feedback on their performance if they were in the mastery quizzing condition.[2] Feedback was always a statement about the correct answer, taken directly from the original passage in the learning phase, regardless of whether the participant answered the question correctly or incorrectly in the mastery quiz condition.

Procedure

Instructions to participants informed them that they were to read several passages about various concepts in psychology with the intention of learning the content for a later memory test. Participants were first asked to read brief paragraphs about ten different concepts in psychology, with a total of ten minutes permitted for reading all ten paragraphs. Each of the ten paragraphs covered unique topics with no overlap between paragraphs. Next, participants engaged in practice quizzes based on the condition to which they had been randomly assigned (i.e., one attempt, two attempts, three attempts, mastery). Participants could advance through the survey at their own pace and had the option to leave questions blank. Quizzes were presented in the same order, but each quiz had its questions randomized. For example, across conditions, all participants would see the same quiz for Quiz 1, but the order of questions and answers within that quiz would be randomized. After completing their practice quiz or quizzes, participants engaged in a brief math distractor task. This distractor task was composed of five one-minute math problems. The final test format matched the quiz format such that each four-option multiple choice question appeared in a random order and participants could advance at their own pace. The final test included only new questions that had not appeared on the practice quizzes, though

[2] Note that Experiments 2 and 3 included feedback for the mastery quizzing condition and the other quizzing conditions, as well.

it tested the same concepts that had appeared both in the learning phase and in the quizzing phase.

Results

Means and standard deviations of the conditions for each condition can be viewed in Table 1. To account for a violation of the analysis of variance (ANOVA) assumption of equal variability across conditions, I ran a test of homogeneity of variances, which was not statistically significant (i.e., $p = .119$).

A one-way ANOVA revealed a significant effect of condition, $F(3, 156) = 11.91$, $p < .001$, $\eta^2 = .186$. Tukey honestly significant difference (HSD) tests revealed that the mastery quizzing condition was significantly different from each of the standard quizzing techniques (i.e., one quiz vs. mastery, $p = .004$; two quizzes vs. mastery & three quizzes vs. mastery, $p < .001$), but none of the standard quizzing conditions were significantly different from each other ($p > .05$). However, effect sizes revealed that some of the differences between conditions were more substantial than others. Table 2 lists confidence intervals, Hedge's g, and p-values for each of the six comparisons from the Tukey test. All statistically significant comparisons also had large effect sizes, while the others had small to medium effect sizes. I will further extrapolate on the implications of these effect sizes later in the paper.

Next, I split participants' data in the mastery condition to determine whether there were statistically significant differences between those who were in the standard quizzing conditions and their counterparts in the mastery condition when controlling for number of quizzing attempts. Since this variable was determined by the participants themselves, I was not able to control sample size for those coming from the mastery condition for these specific comparisons. For each of the following independent samples t-tests I ran Levene's test for equality of

variances to check the assumption of equality of variances. When that assumption was violated, I accounted for this by using the respective degrees of freedom, t-values, and p-values associated with equal variances not assumed.

Although the means show a numerical advantage for those in the mastery condition who only needed one quiz attempt (n = 2, M = 8.00) and those in the one quiz condition (n = 39, M = 7.49), there was not a statistical difference, $t(39) = -4.07$, $p = .686$, $g = .295$. Note that this lack of a statistical difference is likely due, in part, to the small sample of individuals in the mastery condition who only needed one attempt. For those who required two quiz attempts in the mastery condition (n = 13, M = 9.46), there was a significant difference when comparing their final test performance to those in the two-quiz condition (n = 40, M = 6.58), $t(33.065) = -6.84$, $p < .001$, $g = 1.74$. For those who required three attempts in the mastery condition (n = 5, M = 9.60), there was a significant difference when comparing their final test performance to those in the three-quiz condition (n = 38, M = 7.11), $t(27.087) = -5.673$, $p < .001$, $g = 1.16$. Of the remaining participants in the mastery condition, 17 required four to eight quiz attempts to achieve mastery, and six never achieved mastery in the quizzing phase and were forced by the Qualtrics program to progress onto the next phase of the experiment after taking nine mastery quizzes. Of those who did achieve mastery (n = 38), the correlation between number of quiz attempts and final performance was not statistically significant, $r = .224$, $p = .176$.

Discussion

Experiment 1 indicated that there was a benefit for the mastery quizzing condition relative to all other quizzing conditions. The goal of Experiment 1 was to determine if mastery quizzing provided an additional benefit beyond standard quizzing methods. Post hoc comparisons and additional analyses revealed that mastery quizzing was significantly better than

one, two, or three standard quizzes. In addition, when controlling for number of quiz attempts, there was still a benefit for mastery quizzing in the two and three quiz comparisons.

When comparing the standard quizzing conditions, and excluding the mastery condition, there was no significant benefit related to number of quiz attempts. This contradicts research investigating the benefits of frequent testing (Roediger & Karpicke, 2006b) and spacing (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). A meta-analysis on the spacing effect revealed a robust benefit of spacing for intervals shorter than one day (Cepeda et al., 2006), but it is important to point out that while the present study explored differences in number of quiz attempts, there was not a direct comparison of spaced to massed practice. Additionally, there is literature that explores potential effects of different retrieval schedules that may illuminate the present findings.

Karpicke and Roediger (2010) explored the issue of differing retrieval schedules (i.e., fixed vs. expanding retrieval schedules) with text learning. The basic premise of expanding schedules is to increase the spacing interval between successive retrieval attempts, whereas fixed schedules keep the spacing consistent between successive retrieval attempts. Surprisingly, there was not a great deal of research on how spacing affects text learning. They sought to develop a better understanding of how the effects from spacing and different retrieval schedules would affect educationally relevant materials by manipulating how many four-minute intervening periods took place between initial study and test. Overall, there was a significant testing effect found in their research. However, there was no significant difference between performance on expanding schedules and regular schedules. Therefore, spacing schedules did not appear to have a significant impact on long-term retention. These examples may explain why there was not a steady increase in final test performance as number of quizzes increased.

More central to the current paper, the significant benefit of mastery quizzing might support the notion that including feedback would have a major effect on final test performance. That is, given that only the mastery quizzing condition had access to feedback, the benefits shown might only reflect this difference. Experiment 2 thus sought to further explore the benefits of feedback in the mastery condition by including feedback in the standard quizzing conditions.

CHAPTER 3: EXPERIMENT 2

Experiment 2 examined the extent to which the benefit of multiple quiz attempts stems from learners' access to feedback, for both mastery and standard quizzing methods. To accomplish this, Experiment 2 was largely identical to Experiment 1 with the exception that each condition had access to feedback. As with the prior experiment, this allows for a comparison of mastery quizzing to standard quizzing, but this time with feedback provided in the latter case. It also allows for a comparison of one versus multiple quiz attempts in the standard quizzing condition (now with feedback), versus the mastery quizzing condition.

I hypothesized that mastery quizzing would lead to superior performance even when all conditions had feedback, and that this would be reflected in the amount of time spent looking at feedback. I also hypothesized that participants in the mastery condition would spend more time attending to feedback, which would in turn manifest in their increased memory performance.

Method

Participants

After conducting a power analysis using G*Power software (Faul et al., 2007), the required sample size for detecting a medium effect (i.e., $d = .25$) with .8 power was 159 participants. Participants were recruited using Amazon Mechanical Turk and were compensated for their time. Due to how quickly participants signed up for Experiment 1, I decided to lower the compensation to $3.00 for a 40-minute study. I collected the same demographic data outlined in Experiment 1. Participants were prohibited from signing up for this study if they were in Experiment 1. Due to an error in programming, I had to re-run the mastery quizzing condition

and decided to increase the time made available to participants to 50 minutes, thereby increasing the incentive to $5.00.

The initial random assignment of participants to conditions led to there being 59 participants in the one quiz condition and 56 participants in the two-quiz condition. Due to the programming error, 57 participants ended up in the mastery condition, leading to a total of 172 participants. Attrition did not differ greatly among the three conditions.

Design

This experiment was a one factor design with three levels: one quiz attempt, two quiz attempts, and mastery quizzing (i.e., with as many quiz attempts as are necessary to achieve a perfect score). Since the findings from Experiment 1 revealed no significant difference among the three standard quizzing conditions, the three-quiz condition was dropped from Experiment 2.

Materials

Materials for Experiment 2 were identical to those used in Experiment 1.

Procedure

The procedure from Experiment 2 was identical to Experiment 1 with the exception that feedback was now provided for each of the practice quizzes in the standard quizzing conditions. After taking a practice quiz, participants were shown the correct answer for each question, along with brief descriptive feedback (see example in Experiment 1). Participants viewed feedback after answering a quiz question; therefore, feedback was always immediate. This feedback was presented without a time limit. All other methods were consistent with the procedure from Experiment 1.

Results

Descriptive statistics for the conditions can be viewed in Table 3. Levene's test of homogeneity of variances was significant ($p = .002$), indicating that the distribution I violated the assumption of homogeneity of variances. Accordingly, a Welch's ANOVA was conducted, which revealed a significant difference among the three quizzing conditions, $F(2, 110.712) = 9.58$, $p < .001$, $\omega^2 = .09$.

I ran a Games-Howell post hoc test to account for the violation of the assumption of homogeneity of variances. Results from the Games-Howell test revealed a significant difference between the one quiz condition and mastery condition ($p < .001$), a marginally significant difference between the two-quiz condition and mastery condition ($p = .055$), and no significant difference between the one quiz condition and two quiz condition ($p = .20$). While the one quiz versus mastery comparison had a medium to large effect size ($g = .79$), both other comparisons had small to medium effect sizes. Table 4 displays the effect sizes and p-values for each comparison from the Games-Howell test.

As in Experiment 1, I split participants' data in the mastery condition to determine whether there were statistically significant differences between those who were in the standard quizzing conditions and their counterparts in the mastery condition when controlling for number of quizzing attempts. There was no significant difference between those in the one quiz condition ($n = 59$, $M = 7.27$) and those in the mastery condition who needed one quiz attempt to move out of the quizzing phase of the experiment ($n = 2$, $M = 9.00$), $t(59) = -.929$, $p = .357$, $g = .67$. However, the small sample size for the subset of participants in the mastery condition makes this finding difficult, if not impossible, to present as tenable. There was a significant difference between those in the two-quiz condition ($n = 56$, $M = 8.11$) and those in the mastery condition

who needed two quiz attempts (n = 20, M = 9.55), t(73.442) = -3.51, p = .001, g = .63. Of the remaining participants in the mastery condition, 30 required four to nine quiz attempts to achieve mastery, and five never achieved mastery in the quizzing phase and progressed to the next phase of the experiment after taking nine mastery quizzes. Of those who did achieve mastery (n = 52), the correlation between number of quiz attempts and final performance was not statistically significant, r = .018, p = .901.

Discussion

Consistent with Experiment 1, Experiment 2 revealed an overall benefit of mastery quizzing relative to standard quizzing methods even when feedback was provided for all conditions. Thus, the beneficial effect of mastery quizzing persisted even when the standard quizzing conditions also included correct answer feedback. Of note, relative to Experiment 1, the overall effect size was somewhat smaller.

With feedback now normalized across conditions, one of the remaining differences between mastery and the standard quizzing conditions is the implied need to attend to and process the feedback given to learners. Therefore, Experiment 3 sought to further explore this explanation by including a condition that forced participants to rate their level of understanding of the concepts they were learning about as they progressed through the quizzes. This would potentially serve as an experimentally feasible comparison to what students would need to do to feel comfortable with information on their quizzes as they prepared to take high-stakes exams. This novel manipulation required learners to check in with their own knowledge by asking them to rate their level of understanding on the concepts in the quiz, therefore providing an alternative way of addressing learners' attention to what they are learning in a non-mastery condition.

CHAPTER 4: EXPERIMENT 3

Experiment 3 was designed to extend the findings from Experiment 2 by further exploring whether the benefits from mastery quizzing stem from attending to and processing the feedback given. A mastery quizzing condition was compared to a standard quizzing condition with feedback, as well as a standard quizzing condition in which a rating task was included to encourage processing of the feedback provided. One possibility is that the mastery quizzing method encourages extensive processing of feedback to achieve the required perfect score. Research on problem solving suggests that individuals may not transfer learning from one context to a novel, related context unless they are explicitly given a hint to do so (Perfetto, Bransford, & Franks, 1983). By including a rating task in a standard quizzing procedure, to encourage attention to the feedback given, I investigated if standard quizzing can produce final test performance that approaches that of mastery quizzing. The logic here is to include a condition that simulates a situation in which learners are required to attend to the feedback they are given, even though they are not in the mastery condition (i.e., one where they will be forced to take additional quizzes if they did not sufficiently understand the questions).

Method

Participants

After conducting a power analysis using G*Power software (Faul et al., 2007), the required sample size for detecting a medium to strong effect (i.e., $d = .25$) with .8 power was 158 participants. Participants were recruited using Amazon Mechanical Turk and were compensated for their time. As in Experiment 2, I paid participants $3.00 for participating, and I used the same

demographic questions from Experiments 1 and 2. Participants were excluded if they participated in either Experiment 1 or Experiment 2. Due to an error in programming, I had to re-run the mastery quizzing condition and decided to increase the time made available to participants to 50 minutes, thereby increasing the incentive to $5.00.

Twenty participants did not complete the experiment. Of these 20 individuals, seven were from the one quiz condition, seven were from the rating condition, and six were from the mastery condition. The random assignment of participants to conditions led to there being 55 participants in the one quiz condition, 55 participants in the rating condition, and 54 participants in the mastery condition, leading to a total of 164 participants. This distribution does not reflect the pattern of participant attrition that can be found in experiments that use online samples, as attrition did not differ greatly among the three conditions (Zhou & Fishbach, 2016).

Design

A one-factor experiment with three levels: one quiz with detailed feedback given; one quiz with detailed feedback for set amount of time, followed by rating question about participants' subjective judgment of their own knowledge of the concept; and a mastery quiz (i.e., allowing as many quiz attempts as are necessary to achieve a perfect score).

Materials

Materials for Experiment 3 matched those used in Experiments 1 and 2. In the condition where participants were asked to rate their own level of understanding of a given concept they responded to the following, "On a scale of 1 - 5 with 1 being "very poor" and 5 being "very good" rate your knowledge of this concept."

Procedure

The procedure for Experiment 3 was largely identical to Experiment 2. The only exception involved participants in the rating condition who were asked to rate their level of understanding on a given concept. In this condition, participants were shown their feedback after completing each item during the practice quiz, with the additional rating question, “On a scale of 1 - 5 with 1 being "very poor" and 5 being "very good" rate your knowledge of this concept.” Participants could not advance until they had responded to this question.

Results

A Levene’s test of homogeneity of variances indicated that the assumption of homogeneity of variances was violated ($p < .001$). A subsequent Welch’s ANOVA revealed a significant difference among the three quizzing conditions, $F(2, 85.024) = 42.50$, $p < .001$, $\omega^2 = .33$. Descriptive statistics for the conditions can be viewed in Table 5.

A Games-Howell post hoc test was used to account for the violation of the assumption of homogeneity of variances. Results from the Games-Howell test revealed a significant difference between the one quiz condition and mastery condition ($p < .001$, $g = 1.13$), a statistically significant difference between the one quiz with a rating task condition and mastery condition ($p < .001$, $g = 1.46$), and no significant difference between the one quiz condition and the rating condition ($p = .49$, $g = .22$). Table 6 outlines confidence intervals, effect sizes, and p-values for comparisons in Experiment 3.

Because both the one quiz with feedback and one quiz with a rating task conditions only allowed for participants to have one quiz, I ran a subsequent one-way ANOVA comparing these conditions to those in the mastery condition who moved on from the quizzing phase after one attempt (i.e., they achieved mastery after one quiz). This group included seven participants ($M =$

9.57). Forty-five participants in the mastery condition achieved mastery in two to nine attempts, and five participants never achieved mastery within the nine quiz attempts included in the experiment. A statistically significant Levene's test ($p = .02$) meant that I had violated the assumption of homogeneity of variances. A Welch's ANOVA revealed a statistically significant difference among these three groups, $F(2, 36.845) = 23.53$, $p < .001$, $\omega^2 = .28$. Results from the subsequent Games-Howell test revealed a significant difference between the one quiz condition and mastery condition ($p < .001$, $g = .89$) and a statistically significant difference between the one quiz with a rating task condition and mastery condition ($p < .001$, $g = 1.15$). While the effect sizes for the two comparisons involving the mastery condition decreased when controlling for number of attempts, they remained large. Of those who did achieve mastery ($n = 52$), the correlation between number of quiz attempts and final performance was not statistically significant, $r = -.119$, $p = .401$.

Discussion

Based on the findings from Experiment 2, Experiment 3 was primarily designed to determine if the mastery condition's benefit arose from participants' need to attend to and process the feedback they were given. Accordingly, a condition was included with participants required to rate their understanding of each concept. In contrast to my original hypothesis that this condition would match the benefits of mastery when controlling for number of quiz attempts, this experiment revealed a significant benefit of mastery relative to both one-quiz conditions, even when controlling for number of quiz attempts. This suggests that there is an additional benefit for mastery quizzing beyond the presence of feedback and beyond the need to attend to and process the feedback.

CHAPTER 5: GENERAL DISCUSSION

Review of Present Findings

In three experiments, I explored the potential memory benefits of a mastery quizzing technique in which learners were required to complete quizzes until they achieve mastery (i.e., a perfect score). Experiment 1 compared taking one, two, or three quizzes to the mastery condition in which participants were required to take quizzes until they could achieve a perfect score. Final test performance was significantly higher for the mastery condition relative to the other three conditions. Since I considered this could be due to the role of feedback, Experiment 2 included feedback in both the standard quizzing conditions and the mastery condition. I compared one quiz to two quizzes to mastery, and again discovered a significant memory benefit in the mastery condition when considering final test performance. Lastly, Experiment 3 sought to investigate whether the benefits from mastery quizzing stem from attending to and processing the feedback given. A mastery quizzing condition was compared to a standard quizzing condition with feedback, as well as a standard quizzing condition in which a rating task was included to encourage processing of the feedback provided. Once again, mastery led to better overall performance on the final test. My results support the conclusion that mastery quizzing benefits long-term retention over and beyond standard quizzing techniques, even when controlling for various factors such as the number of quiz attempts, the presence of correct answer feedback, and the need to attend to and process feedback.

Account of Benefits

Past studies that have explored related topics have demonstrated that there are benefits associated with several of the aspects of the mastery model (e.g., frequent testing, feedback, spacing). One possible explanation for the overall benefit of mastery is that it recruits several active learning strategies, such as the examples mentioned previously, and this leads to better long-term retention.

Regarding frequent testing, research supports the notion that frequent attempts at retrieval can lead to better long-term retention (Roediger & Karpicke, 2006b). While this may be one of the driving forces behind the mastery benefit, Experiments 1 and 2 demonstrate that it cannot fully explain this benefit. When controlling for number of quiz attempts, mastery still led to higher final test performance, therefore other factors must be taken into consideration.

The role of feedback was also proposed as an important contributor to the benefit of mastery quizzing but Experiments 2 and 3 demonstrate that when normalizing feedback across conditions there is still a significant benefit of mastery quizzing for final memory performance. The explanation for this benefit will likely not be attributable to one factor, so perhaps it is more productive to consider how these factors combined can contribute to an additive memory benefit over and beyond each effect in isolation. Future research could consider the use of a yoked design where participants are joined together with one or more conditions matched in order to better isolate the effects of specific variables.

There is considerable research on the positive impact that retrieval effort has on long-term memory (Carpenter & DeLosh, 2006; Bjork, 1999, Carpenter, 2011). The mastery model may recruit these mechanisms by presenting a scenario to the learner that requires more effort than standard quizzing techniques. It is important to point out that, according to this study's

findings, this need for increased effort would be implied rather than directly stated. Regardless of which condition participants were in, they were told initially that they may be in a situation where they would see similar concepts again, meaning that participants were never directly told that they would be forced to achieve perfection to progress through the experiment. This suggests that any benefit of mastery condition is unrelated to participants' motivation to achieve perfection from explicit awareness, and instead hints at a different potential benefit.

In a recent meta-analysis, Chan et al. (2018) suggested that the benefits of retrieval practice can be most attributed to two clusters of theories: resource theories and integration theories. Resource theories suggest that retrieval practice reduces PI, contributing to better performance on a final test, or that it aids attention in restoring any capacity that was depleted during the original learning of the material. Most typically, interference is thought to happen during retrieval, leading to retrieval competition (Kuhl & Wagner, 2009). The extent to which interference occurs is related to the number and strength of irrelevant memories, which can be compared to research on divided attention, which shows that participants attempt to exert their attentional control over whichever task(s) they consider to be most in need of their attention (Gopher, 1993). Perhaps retrieval practice helps allay the effects of these competing stimuli, therefore leading to enhanced memory performance. The benefits of mastery quizzing could be that they increase the accessibility of the originally learned information in memory, if we consider this from a resource perspective.

Some research on metacognition can supplement these ideas as well. Some recent research has looked at learners' perceptions of stability bias. Stability bias refers to learners' belief that the current accessibility of information will remain relatively stable over time (Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011). What some research shows is that this

can alter retrieval fluency, and not always for the better (Soderstrom & Bjork, 2015). Therefore, illusions of accessibility can harm retrieval in a way that may lead individuals to choose less effective learning techniques (Kornell, 2009; Dunlosky & Rawson, 2015). However, forcing learners to continue to engage in retrieval practice until they achieve a perfect score may help avoid these pitfalls.

Integration theories emphasize the importance of associations between bits of information. These theories share some parallels with the resource account in that they emphasize the importance of accessibility of the originally learned information, but the important difference is that the practice test leads to covert retrieval of the original information, which binds the original information to whatever new information the learner is getting (e.g., associations, context). This ultimately improves conceptual organization of the information, which also promotes long-term memory performance. This account leans more heavily on semantic-based explanations of the testing effect such as the mediator effectiveness hypothesis (Carpenter, 2011). In the context of the current study, mastery quizzing may be leading to better integration of information relative to standard quizzing techniques.

Episodic context accounts of the testing effect state that retrieval practice encourages the learner to reinstate the original learning context, leading to better memory on a final test (Whiffen & Karpicke, 2017). One could postulate that this contributed to the increased final test performance since items in each phase of the experiment were presented in a similar order. However, since this was true of all conditions (i.e., mastery and non-mastery), it is unlikely that this uniquely benefitted those in the mastery condition. If all conditions were reinstating their original learning context when performing on the final test, it could also be the case that this

strategy enhanced the memory benefit in the mastery condition greater than in the other conditions.

Limitations

It is important to consider potential shortcomings or limitations that may have impacted the study's findings. It could be the case that individual differences drove the mastery benefit when controlling for number of attempts. Considering the findings in Experiment 2, specifically, all conditions were essentially matched when controlling for number of quiz attempts (i.e., all participants saw the same quizzes and had the same feedback). Since the sample sizes were small for those in the mastery condition that ended up in these comparisons, one could argue that the benefit was driven by individuals who were already high performers. There is research to suggest that those who are high-knowledge learners may benefit more when they are required to engage in more effortful processing of the information (e.g., coming up with their own examples, establishing connections that are merely inferred). A recent set of experiments explored this idea by investigating how students of differing ability levels would perform on different retrieval strategies, but also sought to better understand the metacognitive calibration of these students that were either deemed high performers or low performers (Carpenter et al., 2016). Therefore, there is certainly a possibility one must consider in which the mastery benefit is driven by those who already have the potential to score higher on measures of memory. Minear, Coane, Boland, Cooney, and Albat (2018) demonstrated that the magnitude of the testing effect differed depending on item difficulty and participants' ability levels (i.e., those with higher fluid intelligence demonstrated a stronger testing effect on difficult items, whereas those with lower fluid intelligence demonstrated a stronger testing effect on easy items). Both studies suggest that

individual differences should be taken into consideration when considering the benefits of retrieval practice.

One noteworthy methodological limitation in the present study is the error in programming that led to the re-running of the mastery condition for all three experiments. Without random assignment to conditions, the internal validity of the study is certainly in question. Furthermore, after deciding to increase the time and incentive for the mastery condition based on frequent feedback that 40 minutes was not enough, this may have altered my findings. While there is research demonstrating that simply motivating participants with money does not significantly impact their memory performance (Nilsson, 1987), it is worth pointing out that I may have unintentionally recruited more studious participants in the mastery condition by having a greater incentive.

Another important limitation to point out is how small the sample sizes were in the mastery conditions when controlling from number of quiz attempts. While these were still noteworthy comparisons, their generalizability is harmed when we are operating with sample sizes as small as two. Future research would benefit from re-running these comparisons with adequate sample sizes. Additionally, it is important to point out that the retention interval for this set of experiments was five minutes, meaning it is not clear whether these differences would hold up after longer delays.

Implications and Future Research

The benefits of mastery quizzing offer interesting theoretical implications for the study of the testing effect. If memory for information was better when engaging in this technique, this could extend our understanding of how specific factors (i.e., number of retrieval attempts, feedback) affect the process of retrieval. Mastery quizzing may also bring together several

empirically tested factors that seem to promote learning (e.g., elaboration, spacing) and reveal that, when combined, these different components may produce an even greater memory benefit than they could in isolation.

In focusing on practical implications, the findings from the present study imply that mastery quizzing is a more beneficial learning tool than standard quizzing, even when controlling for factors such as the presence of feedback and the number of quiz opportunities. This technique is a utilizable tool that educators could begin implementing in their classes in order to help their students learn better.

To simulate more realistic classroom environments in which students have several days in-between their quizzes and tests, future research should seek to explore whether the mastery model can benefit long-term retention at longer intervals. One of the biggest challenges with drawing conclusions about student study behavior from laboratory studies on memory is that the conditions necessary to infer causation harm the external validity of the study, leading to potentially artificial scenarios that do not directly translate into classroom relevant situations. Therefore, this research would benefit from a classroom study that directly tests the efficacy of the mastery model.

With online, unsupervised quizzing of this nature, an instructor might be concerned that students may search for answers from their written materials or outside resources, rather than learn and retrieve the information from memory. One might be concerned that, worse yet, students will work on the quizzes with classmates, get information about the questions from classmates, or even get the answers themselves from classmates. I emphasize, however, that these quizzes are intended to serve as learning tools rather than assessment tools, and some useful learning should occur even if students are consulting written materials or interacting with

fellow students. Moreover, given that these online quizzes are designed to serve as learning tools for students' benefit rather than assessment tools for assigning students' grades, instructors typically set up their courses so that the quizzes contribute a relatively small percentage toward students' grades, such that course grades are not unduly influenced by the quiz performance.

Summary and Conclusion

In conclusion, this paper offers preliminary evidence to support the use of a mastery model to promote effective learning in the classroom. Over and beyond standard quizzing techniques in which students are permitted one or more attempts and are scored based on their best performance, the mastery model forces students to continue to complete the quiz until they achieve a perfect score. If educators want to increase their students' learning, they should seek to employ a mastery quizzing technique in their own classes.

CHAPTER 6: TABLES

Table 1

Descriptive Statistics by Quiz Group for Experiment 1

	M	SD	SE	N
One	7.49	1.70	.27	39
Two	6.58	1.80	.28	40
Three	7.11	2.25	.37	38
Mastery	9.12	1.55	.23	44

Table 2

Effect Sizes (g), P-values, and 95% Confidence Intervals (CI) by Quiz Group Comparison for Experiment 1

	g	p	CI95
One vs. Two	.52	.151	-.20, 2.03
One vs. Three	.19	.817	-.75, 1.51
One vs. Mastery	-1.0	.004**	-2.56, -.37
Two vs. Three	-.26	.612	-1.65, .59
Two vs. Mastery	-1.50	.000***	-3.47, -1.29
Three vs. Mastery	-1.04	.000***	-2.95, -.74

**Statistically significant at level $p < .01$

***Statistically significant at level $p < .001$

Table 3

Descriptive Statistics by Quiz Group for Experiment 2

	M	SD	SE	N
One	7.27	2.61	.34	59
Two	8.11	2.58	.35	56
Mastery	9.12	1.98	.26	57

Table 4

Effect Sizes (g), P-values, and 95% Confidence Intervals by Quiz Group Comparison for Experiment 2

	g	p	CI95
One vs. Two	-.32	.20	-1.99, .31
One vs. Mastery	-.79	.000***	-2.87, -.83
Two vs. Mastery	-.44	.055	-2.05, .02

***Statistically significant at level $p < .001$

Table 5

Descriptive Statistics by Quiz Group for Experiment 3

	M	SD	SE	n
One	7.18	2.78	.37	55
Rating	6.58	2.69	.36	55
Mastery	7.80	2.52	.34	54

Table 6

Effect Sizes (f), P-values, and 95% Confidence Intervals (CI) by Quiz Group Comparison for Experiment 3

	g	p	CI95
One vs. Rating	.22	.49	-.64, 1.84
One vs. Mastery	-.23	.000***	-3.27, -1.38
Rating vs. Mastery	-.46	.000***	-3.84, -2.01

***Statistically significant at level $p < .001$

REFERENCES

Agarwal, P. K., D'Antonio, L., Roediger III, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131-139.

Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy?. *Acta Psychologica*, *128*(1), 110-118.

Barnett, S. M. & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and Performance* (pp. 435-459). Cambridge, MA: MIT Press.

Butler, A. C. & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616.

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

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547-1552.

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., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*(2), 353-375.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*(6), 760-771.

Carpenter, S. K. & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92*, 128-141.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380.

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.

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*(11), 1111-1146. doi:10.1037/bul0000166

Chan, J. C. K., McDermott, K. B. & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially non-tested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553-571.

Craik, F. I. M. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294.

Darley, C. F. & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91*(1), 66-73.

Dunlosky, J. & Lipko, A. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science. 16*, 228-232.

Dunlosky, J. & Rawson, K. A. (2015). Do students use testing and feedback while learning? A focus on key concept definitions and learning to criterion. *Learning and Instruction, 39*, 32-44.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, *39*, 175-191. doi:10.3758/ BF03193146

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology, 28*(1), 135-142.

Gopher, D. (1993). The skill of attention control: Acquisition and execution of attention strategies. In *Skill of Attention Control* (pp. 299-322).

Hartwig, M. K. & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126-134.

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, *17*(6), 797-801.

Hughes, A. D. & Whittlesea, B. W. A. (2003). Long-term semantic transfer: An overlapping-operations account. *Memory & Cognition, 31*(3), 401-411.

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology, 35*(4), 513-521.

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

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61,* 237-284.

Karpicke, J. D. & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38*(1), 116-124.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297-1317.

Kornell, N. & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219-224.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449-468.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787-794.

Kuhl, B. A. & Wagner, A. D. (2009). Forgetting and retrieval. In G. G. Berntson, J. T. Cacioppo (Eds.). *Handbook of neurosciences for the behavioral sciences* (pp. 1-20). John Wiley and Sons.

Kulik, J. A. & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 1*, 79-97.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210-212.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 494-513.

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

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

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*(1), 18-26.

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474-1484.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory, 24*(2), 257-271. doi:10.1080/09658211.2014.1001992

Mulligan, N. W. & Lozito, J. P. (2004). Self-generation and memory. *Psychology of Learning and Motivation, 45*, 175-214.

Myers, D. G. & DeWall, C. N. (2016). *Exploring psychology in modules* (10th ed.). New York, NY: Worth Publishers.

Nelson, D. L., McEvoy, C. L., & Schrieber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments & Computers, 36*, 402-407.

Nilsson, L.-G. (1987). Motivated memory: Dissociation between performance data and subjective reports. *Psychological Research, 49*(2-3), 183-188. doi:10.1007/BF00308685

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.

Perfetto, G. A., Bransford, J. D., & Franks, J. J. (1983). Constraints on access in a problem solving context. *Memory & Cognition, 11*(1), 24-31.

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*(4), 437-447.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335. doi:10.1126/science.1191465

Pyc, M. A. & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 737-746. doi:10.1037/a0026166

Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: When and why are judgments accurate? *Memory & Cognition, 42*, 886-897.

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

Rhodes, M. G. & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131-148. doi:10.1037/a0021705

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382-395.

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. doi:10.1016/j.tics.2010.09.003

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

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

Roediger, H. L. & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: Essays in honor of Robert A. Bjork* (pp. 23–48). New York, NY: Psychology Press.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation, 55*, 1-36. doi:10.1016/B978-0-12-387691-1.00001-6

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin,* 1-31.

Rowland, C. A. & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review.*

Salomon, G. & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist, 24*(2), 113-142.

Soderstrom, N. C. & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176-199.

Whiffen, J. W. & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1036-1046. doi:10.1037/xlm0000739

Wulf, G. & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review, 9*(2), 185-211.

Zhou, H. & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*. doi:10.1037/pspa0000056

APPENDICES

Appendix A

Passages from Learning Phase of Experiments 1 – 3

1. The cerebral cortex is the body's ultimate control and information-processing center. The cerebral cortex can be divided into four pairs of lobes: the frontal lobes, the parietal lobes, the occipital lobes, and the temporal lobes. The frontal lobes are involved in judgment and making plans. The parietal lobes are involved in receiving sensory input for touch and body position. The occipital lobes are involved in receiving information from our visual fields. The temporal lobes are involved in processing auditory information by receiving sensory input from the ears.
2. In operant conditioning, reinforcement is any event that strengthens the behavior it follows. Positive reinforcement increases behaviors by presenting something desirable. Negative reinforcement increases behaviors by stopping or reducing something undesirable. Punishment is an event that tends to decrease the behavior that it follows. Positive punishment decreases behaviors by presenting something undesirable. Negative punishment decreases behaviors by stopping or reducing something desirable.
3. Atkinson and Shiffrin (1968) proposed a Modal model of memory with three stages. We first record to-be-remembered information as a fleeting sensory memory. Sensory memory is the immediate, very brief recording of sensory information in the memory system. From there, we process information into short-term memory, where we encode it through rehearsal. Short-term memory is activated memory that holds a few items briefly,

such as the seven digits of a phone number while calling, before the information is stored or forgotten. Finally, information moves into long-term memory for later retrieval. Long-term memory is the relatively permanent and limitless storehouse of the memory system. This includes knowledge, skills, and experiences.

4. The Big Five personality factors are conscientiousness, agreeableness, neuroticism, openness, and extraversion. These dimensions may be objectively measured, they are relatively stable over the life span, and they apply to all cultures in which they have been studied. A very conscientious person would be very organized, careful, and disciplined. A very agreeable person would be soft-hearted, trusted, and helpful. A very neurotic person would be anxious, insecure, and self-pitying. A very open person would be imaginative, independent, and prefer variety. A very extraverted person would be sociable, fun-loving, and affectionate.
5. Group behavior can impact our own behavior in a variety of circumstances. The bystander effect demonstrates that when many other people are present, the chances that an individual will help someone in need decreases. Social loafing is the tendency for people in a group to exert less effort when pooling their efforts toward attaining a common goal than when individually accountable. Group polarization is the enhancement of a group's prevailing inclinations through discussion within the group. Groupthink is the mode of thinking that occurs when the desire for harmony in a decision-making group overrides a realistic appraisal of alternatives.
6. Jean Piaget proposed that children develop cognitively through a series of stages. In the sensorimotor stage (birth – 2 years old), infants know the world mostly in terms of their sensory impressions and motor activities. In the preoperational stage (2 – 6/7 years old),

children learn to use language but do not yet comprehend the mental operations of concrete logic. During the concrete operational stage (7 – 11 years old), children gain the mental operations that allow them to think logically about concrete events. During the formal operational stage (12 years old through adulthood), people begin to think logically about abstract concepts.

7. Descriptive research methods are used to observe and record behavior using case studies, naturalistic observations, or surveys. Correlational research methods are used to detect naturally occurring relationships, and to assess how well one variable predicts another. Experimental research methods are used to explore cause and effect relationships between factors by manipulating an independent variable (i.e., the factor whose effect is being studied).
8. A neuron, or nerve cell, is the basic building block of the nervous system. A neuron contains dendrites, which are branching extensions that receive messages. Dendrites conduct neural impulses toward the cell body of the neuron. From there, the neural impulse travels down the axon, which passes messages away from the cell body through its terminal branches. The terminal branches of an axon form junctions with other cells and pass the neural impulse on to other neurons or to muscles or glands.
9. Depth perception enables us to estimate an object's distance from us. We often depend on certain monocular cues (i.e., depth cues that only require one eye) when judging an object's distance from us. The cue of relative height refers to our tendency to perceive objects higher in our field of vision as farther away. Relative size refers to our tendency to perceive smaller objects as being farther away. Interposition refers to our tendency to assume that if one object partially blocks our view of another, the object doing the

blocking is closer to us. Linear perspective refers to our tendency to perceive that parallel lines appear to meet in the distance.

10. Anxiety disorders are marked by distressing, persistent anxiety and often dysfunctional anxiety-reducing behaviors. Generalized anxiety disorder is an anxiety disorder in which a person is continually tense, apprehensive, and uneasy. Panic disorder is an anxiety disorder marked by unpredictable, minutes-long episodes of intense dread in which a person experiences terror and accompanying chest pain, choking, or other frightening sensations. A phobia is an anxiety disorder marked by a persistent, irrational fear and avoidance of a specific object, activity, or situation. Obsessive-compulsive disorder is a disorder characterized by unwanted repetitive thoughts (i.e., obsessions), actions (i.e., compulsions), or both.

Appendix B

Quiz and Final Test Questions from Experiments 1 – 3

Quiz 1:

1. Which lobes in the cerebral cortex are responsible for planning and judgment?
    a. Frontal
    b. Temporal
    c. Parietal
    d. Occipital
2. _____ increases behaviors by stopping or reducing something undesirable.
    a. Negative reinforcement
    b. Positive reinforcement
    c. Negative punishment
    d. Positive punishment

3. _____ is activated memory that holds a few items briefly, such as the seven digits of a phone number while calling, before the information is stored or forgotten.
    a. Sensory memory
    b. Short-term memory
    c. Long-term memory
    d. Rehearsal
4. _____ is a personality dimension related to organization, discipline, and care when completing tasks.
    a. Conscientiousness
    b. Agreeableness
    c. Neuroticism
    d. Openness
5. _____ is the enhancement of a group's prevailing inclinations through discussion within the group.
    a. Group polarization
    b. Groupthink
    c. The bystander effect
    d. Social loafing
6. In the _____ stage people begin to think logically about abstract concepts, according to Piaget.
    a. Sensorimotor
    b. Preoperational
    c. Concrete operational

d. Formal operational

7. Which type of research method is used to explore cause and effect relationships between factors?
    a. Descriptive
    b. Correlational
    c. Experimental
    d. Independent variable
8. Which part of a neuron receives messages from other cells?
    a. Dendrites
    b. Cell body
    c. Axon
    d. Terminal branches
9. Which monocular depth cue refers to our tendency to perceive that parallel lines appear to meet in the distance?
    a. Linear perspective
    b. Interposition
    c. Relative height
    d. Relative size
10. Which anxiety disorder is characterized by unpredictable, minutes-long episodes of intense dread in which a person experiences terror and accompanying chest pain, choking, or other frightening sensations?
    a. Generalized anxiety disorder
    b. Panic disorder

c. A phobia

d. Obsessive-compulsive disorder

Quiz 2:

1. The _____ lobes in the cerebral cortex are responsible for planning and judgment.

    a. Frontal

    b. Temporal

    c. Parietal

    d. Occipital

2. Which operant conditioning term refers to something that increases behaviors by stopping or reducing something undesirable?

    a. Negative reinforcement

    b. Positive reinforcement

    c. Negative punishment

    d. Positive punishment

3. Which type of memory holds a few items briefly, such as the seven digits of a phone number while calling, before the information is stored or forgotten?

    a. Sensory memory

    b. Short-term memory

    c. Long-term memory

    d. Rehearsal

4. Which personality dimension is related to organization, discipline, and care when completing tasks?

a. Conscientiousness

b. Agreeableness

c. Neuroticism

d. Openness

5. What is it called when there is an enhancement of a group's prevailing inclinations through discussion within the group?

a. Group polarization

b. Groupthink

c. The bystander effect

d. Social loafing

6. Which of Piaget's stages of cognitive development reflects an ability to think logically about abstract concepts?

a. Sensorimotor

b. Preoperational

c. Concrete operational

d. Formal operational

7. _____ research methods are used to explore cause and effect relationships between factors.

a. Descriptive

b. Correlational

c. Experimental

d. Independent variable

8. The _____ of a neuron receive(s) messages from other cells.

a. Dendrites

b. Cell body

c. Axon

d. Terminal branches

9. _____ refers to our tendency to perceive that parallel lines appear to meet in the distance.

    a. Linear perspective

    b. Interposition

    c. Relative height

    d. Relative size

10. _____ is an anxiety disorder marked by unpredictable, minutes-long episodes of intense dread in which a person experiences terror and accompanying chest pain, choking, or other frightening sensations.

    a. Generalized anxiety disorder

    b. Panic disorder

    c. A phobia

    d. Obsessive-compulsive disorder

Quiz 3:

1. Art has sustained damage to his brain after having a stroke. His wife notices that he now has difficulty making plans. Which lobe of the cerebral cortex was likely damaged during Art's stroke?

    a. Frontal

    b. Temporal

c. Parietal

d. Occipital

2. If you take painkillers to relieve the pain you are experiencing from a headache, you are demonstrating the operant conditioning principle of

    a. Negative reinforcement

    b. Positive reinforcement

    c. Negative punishment

    d. Positive punishment

3. We would have difficulty remembering a series of twelve digits for more than several seconds, but we can remember seven-digit phone numbers when we are about to call someone thanks to the capacity of our

    a. Sensory memory

    b. Short-term memory

    c. Long-term memory

    d. Rehearsal

4. Hannah is very organized, disciplined, and careful when organizing her schedule. Hannah would likely score high on the _____ dimension of the Big Five personality factors.

    a. Conscientiousness

    b. Agreeableness

    c. Neuroticism

    d. Openness

5. When people find themselves in groups of like-minded types, they are especially likely to move to extremes. This is the phenomenon of
    a. Group polarization
    b. Groupthink
    c. The bystander effect
    d. Social loafing
6. Lee can logically understand abstract concepts, such as freedom. Lee is likely in the _____ stage of cognitive development.
    a. Sensorimotor
    b. Preoperational
    c. Concrete operational
    d. Formal operational
7. Dr. White wants to see if caffeine affects exam scores, so he designs a study in which half of his participants drink four cups of coffee before an exam, and the other half drink no coffee before an exam. Given that Dr. White has manipulated the amount of coffee that participants drank, what kind of research method is he using?
    a. Descriptive
    b. Correlational
    c. Experimental
    d. Independent variable
8. A neural message is first picked up by the _____ of a neuron.
    a. Dendrites
    b. Cell body

c. Axon

d. Terminal branches

9. If you stand on train tracks and look towards the horizon they will appear to meet in the distance. Which monocular depth cue does this represent?

    a. Linear perspective

    b. Interposition

    c. Relative height

    d. Relative size

10. Those who experience unpredictable periods of terror and intense dread, accompanied by frightening physical sensations, may be diagnosed with

    a. Generalized anxiety disorder

    b. Panic disorder

    c. A phobia

    d. Obsessive-compulsive disorder

Final Test:

1. Adolescents sometimes have trouble judging whether a choice they make is right or wrong. This could be attributed to the fact that their _____ lobes are not fully developed.

    a. Frontal

    b. Temporal

    c. Parietal

    d. Occipital

2. When you start your car, it beeps at you loudly until you fasten your seat belt. By fastening your seat belt, you end the loud beeping. Which operant conditioning term does this demonstrate?
    a. Negative reinforcement
    b. Positive reinforcement
    c. Negative punishment
    d. Positive punishment
3. You park your car in a public parking lot and are required to enter your license plate information at the meter when paying. While walking to the meter you repeat your license plate number to yourself in your head, so that you can remember what it is when typing the digits in several seconds later. In this example you are storing your license plate information in
    a. Sensory memory
    b. Short-term memory
    c. Long-term memory
    d. Rehearsal
4. Mike is very disorganized, careless, and lacks discipline when completing tasks at work. Mike would likely score low on the _____ dimension of the Big Five personality factors.
    a. Conscientiousness
    b. Agreeableness
    c. Neuroticism
    d. Openness

5. Students who oppose the death penalty meet on their university's campus for a weekly discussion with like-minded peers. After several meetings the group's stance on the death penalty grows more extreme, such that they are all now vehemently opposed to the death penalty rather than just being slightly against it. The change in their behavior is an example of
    a. Group polarization
    b. Groupthink
    c. The bystander effect
    d. Social loafing
6. Gabriella can understand concrete logical problems, but cannot understand abstract logical problems. It is likely that Gabriella has not yet made it to the _____ stage of cognitive development.
    a. Sensorimotor
    b. Preoperational
    c. Concrete operational
    d. Formal operational
7. Dr. Ramos is interested in whether packaging color affects buyers' likelihood of purchasing different flavors of ice cream. She runs a study in which some ice cream brands are packaged in red containers and others are packaged in blue containers to see if one color leads to higher levels of purchasing. What kind of study is Dr. Ramos doing?
    a. Descriptive
    b. Correlational
    c. Experimental

d. Independent variable

8. You accidentally place your hand on a hot surface and immediately pull it away quickly, thanks to the neurons in your hand that alerted you to the pain. Which part of a neuron would have gotten the information that the surface was hot and then passed it on to the rest of your nervous system?
    a. Dendrites
    b. Cell body
    c. Axon
    d. Terminal branches
9. A farmer stares at her parallel rows of corn and sees that they appear to meet in the distance. Which monocular depth cue does this represent?
    a. Linear perspective
    b. Interposition
    c. Relative height
    d. Relative size
10. While driving home from work, Pam suddenly felt intense dread, accompanied by trembling, dizziness, chest pains, and choking sensations. Pam may have
    a. Generalized anxiety disorder
    b. Panic disorder
    c. A phobia
    d. Obsessive-compulsive disorder