2018

Can multiple-choice testing potentiate new learning for text passages? A meta-cognitive approach to understanding the forward testing effect

Sara Dawn Davis
Iowa State University

Can multiple-choice testing potentiate new learning for text passages? A meta-cognitive approach to understanding the forward testing effect

by

Sara D. Davis

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Psychology

Program of Study Committee:
Jason C.K. Chan, Major Professor
Patrick I. Armstrong
Shana K. Carpenter
Michael F. Dahlstrom
Christian A. Meissner

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First, I would like to extend my sincerest gratitude to my major professor, Jason C.K. Chan. I appreciate all of the time and effort you have invested in me, and I am so grateful that I was able to complete my doctoral work under your supervision. I owe a great deal of my professional development to you, and I am a better scientist because of your guidance.

I would also like to thank the other members of my committee for their continued support and guidance over the past three years. It truly takes a village, and I am very thankful for the opportunity to learn from each of you. Thank you for graciously donating your time to me and for being so willing to meet with me, give feedback, and write letters.

I would also be remiss if I didn't thank the strong network of social support that I have. Of course, I need to thank the CCVE members for everything they have done for me throughout my graduate school career, professionally and personally. From proofreading drafts late at night, to answering statistical questions, and encouraging me to run Bayesian analyses even when I didn't want to, your unwavering friendship has been invaluable. I also have to thank Kristin Marshall for all of her support through prelims, teaching, and this dissertation. You are truly a gem, and your positivity has helped me keep my head up through all of these herculean tasks.

Last, I need to thank my partner, Nick, for his unconditional support throughout everything. Your strength is an inspiration to me, and I could not have done this without you. This wasn't always a fun process, but you were there through thick and thin. I simply cannot thank you enough.

ABSTRACT

A burgeoning area of research has begun to examine how retrieval practice can influence future learning that occurs *after* a test. In general, the extant literature has demonstrated a *forward testing effect*, in which prior testing enhances new learning. However, there is not a consensus as to the mechanism that leads to this phenomenon. In the present dissertation, I propose a metacognitive account, in which testing is purported to benefit subsequent learning by leading learners to better attend to and encode material. In particular, individuals who are tested gain valuable information about the nature and difficulty of upcoming tests, which helps guide their strategy use. Under this account, more difficult retrieval (e.g., recall) should lead to a greater metacognitive benefit than easier retrieval (e.g., multiple-choice). In Experiment 1, I compared prior cued-recall, prior multiple-choice, or no prior testing on performance for a criterial test of a text passage. Furthermore, I examined how the match between initial and criterial tests might determine whether testing influences new learning. In fact, prior cued-recall testing enhanced learning to a greater degree than prior multiple-choice testing (relative to no prior testing) regardless of criterial test format. Reading times for each text passage provided preliminary evidence for a metacognitive benefit of testing. Whereas reading times fell across the passages for individuals who were not tested, reading times remained stable in both testing conditions. In Experiments 2 and 3, I aimed to further investigate the metacognitive mechanism underlying the forward testing effect by requiring explicit judgments of learning (JOLs) prior to each testing (or non-testing) episode. Surprisingly, when JOLs were required, there was no forward testing effect observed in Experiment 2 (which included prior cued-recall, multiple-choice, or no-testing) or in Experiment 3 (which included a more difficult prior multiple-choice condition). In both Experiments 2 and 3, reading times remained stable across passages in each

condition, although JOLs indicated less confidence in predicted performance in the tested conditions than in those who were not tested prior to the criterial test.

CHAPTER 1. INTRODUCTION

Despite being the world's economic leader (IMF World Economic Outlook Database, 2016), educational attainment for American students, particularly in Science, Technology, Engineering, and Math (STEM) fields, lags far behind the achievement evident in the rest of the developed world. To illustrate, in the 2015 International Student Assessment, the US ranked 24th in Science and 39th in Mathematics, and this reflects a long-term trend demonstrating American students' underachievement in these fields relative to our international counterparts (National Center for Education Statistics, 2015). In 2012, the President's Council of Advisors on Science and Technology called for an increase in STEM bachelor's degrees of 1 million over projected numbers, but over 40% of students who begin a degree in STEM do not graduate with these qualifications (Chen & Soldner, 2013). As the Council highlighted, success for students in STEM fields depends on the empirical validation of teaching methods used to teach STEM in higher education, and as such, the burden of provision of this evidence falls at least partly upon cognitive scientists studying memory. As researchers who study memory, cognitive psychologists are well positioned to identify techniques that can enhance learning of STEM materials, a necessary building block for encouraging the attainment of STEM degrees. Therefore, an important goal of this dissertation is to identify a technique that can help maximize STEM learning.

In particular, this dissertation focuses on the technique of enhancing new learning with interpolated retrieval practice. In a standard paradigm of this nature, participants or students learn several sections of material, such as lists of words, sections of a text passage, or sections of a video lecture. After all but the final section, learners engage in either interpolated tests (e.g., practicing retrieval for some or all of the material in the prior learning opportunity) or not (e.g.,

learners may review the previous material, engage in an unrelated task, or simply move on to the next new learning opportunity). Importantly, all learners are tested after the final new learning opportunity. A *forward testing effect* has typically been found, such that learners who receive prior testing learn new information better than those who do not receive prior testing. This technique could be instrumental in enhancing STEM education, inasmuch as it encourages long-term learning of material.

While a great deal of research has focused on prior *recall* (e.g., short answer or essay questions), no research has examined how prior *recognition* (e.g., multiple-choice or true/false tests) can enhance new learning. Thus, a primary goal of this dissertation is to examine to what extent multiple-choice testing might enhance learning for STEM materials. By using multiple-choice tests, one can address the question of whether the forward testing effect is due to a metacognitive enhancement, wherein learners gain metacognitive knowledge across testing opportunities. For example, if students take easier tests (such as multiple-choice tests, relative to short answer tests), they may be less likely to adopt more effective study strategies for subsequent learning sessions. Due to their relative ease, I predicted that multiple-choice tests would result in higher confidence in students' learning, which should reduce the likelihood of strategy changes across learning opportunities relative to more difficult short-answer tests. Furthermore, this dissertation also examines how the technique of interpolating retrieval practice during learning can influence performance on a later *cumulative* test. I will return to these predictions later in the Introduction, but will first describe the literature on how retrieval affects old and new learning.

What is the Impact of Retrieval on Old and New Learning?

One of the most effective techniques at enhancing learning that has been identified by research is *retrieval practice.* Retrieval practice can take many forms, including formal high-stakes assessments, low-stakes quizzes, self-testing (such as with flashcards), or covert retrieval (attempting to retrieve a piece of information without providing a response). A wealth of evidence has determined that retrieval practice generally has a large beneficial effect on student learning, such that tested material is retained at a higher rate than material that is simply restudied (Roediger & Karpicke, 2006, Rowland, 2014). In an educational context, testing is an ideal instructional and study tool for students, as it is inexpensive and requires little training to implement. Therefore, it has been widely touted (Dunlosky, Marsh, Nathan, & Willingham, 2013; Pashler, Bain, Bottge, Graesser, Roediger, McDaniel, & Metcalfe, 2007) to the educational community as a useful resource to enhance student achievement.

There is a preponderance of evidence in favor of this *backward testing effect,* the finding that taking a test for some material can enhance the long-term memory for that same material. In contrast, a recent field of study has also begun to examine the *forward testing effect,* or how taking a test can enhance learning that occurs *after* the test. For example, Szpunar, McDermott, and Roediger (2008) required participants to study four word lists (see Figure 1). All participants took a test over the final word list, and performance on this test (from here on referred to as the *criterial test*) served as the critical dependent variable. On the preceding three lists, participants either took a free recall test after each list or not (to equate for the amount of time spent taking a test, participants in the no-test condition completed math problems for the same amount of time). Participants who were previously tested demonstrated superior

performance on the criterial test relative to those who did not take interpolated tests, demonstrating a forward testing effect. This effect has been replicated with related and unrelated word lists (Bauml & Kliegel, 2013; Pastotter, Weber, & Bauml, 2013; Pastotter, Schicker, Niedernhuber, & Bauml, 2011; Szpunar et al.), face-name and face-profession associations (Davis & Chan, 2015; Weinstein, McDermott, & Szpunar, 2011; Yang, Potts, & Shanks, 2017), prose passages (Wissman, & Rawson, & Pyc, 2011, Wissman and Rawson, 2015), and video lectures (Szpunar, Khan & Schacter, 2013; Jing, Szpunar, & Schachter, 2016, Szpunar, Jing, and Schacter, 2014). In fact, enough evidence has amassed in favor of the beneficial forward testing effect that Chan, Meissner, and Davis (2018) conducted a meta-analysis on the existing literature wherein participants were either tested or not before learning new information. An analysis comparing the previously untested conditions and tested conditions on a final test for the last section of materials revealed an effect size of Hedges $g = 0.75$ ($k = 84$), a large effect. This suggests that students would benefit from interpolated testing during the reading of textbook passages, which is of particular interest for this study.

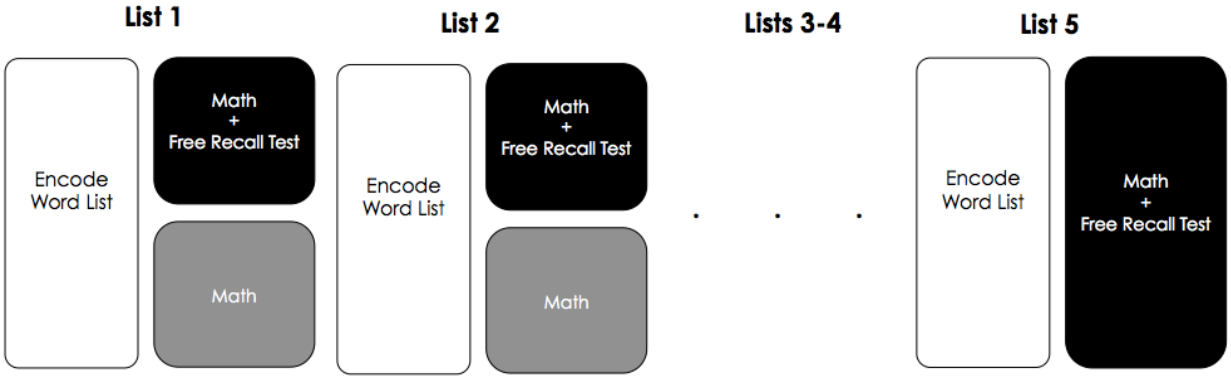

Figure 1. A graphical depiction of the method used by Szpunar et al (2008).

Despite presenting strong evidence in favor of the forward testing effect, this meta-analysis (Chan et al., 2018) also identified several areas that have been under-researched. The most important of these with regard to this dissertation is the test format that is used for the interpolated tests. All of the extant studies have examined the effect of recall, in which participants must engage recollection processes and explicitly generate a response. For example, a recall test might ask participants to recall all of the information they could remember from a previous text passage. However, none of the studies examined whether recognition might also produce a forward testing effect. This is important because a common type of recognition test in educational contexts is multiple-choice testing. In a multiple-choice test, students might be presented with a question (e.g., Who identified the two key components of the laser?) and must select the correct option (e.g., Albert Einstein) out of a series of *foils*, or incorrect distractor items (e.g., Isaac Newton, Gordon Gould, and Stephen Hawking).

Multiple-choice tests are pervasive in the classroom for several reasons. They are favored by students (Zeidner, 1987; Struyven, Dochy, & Janssens, 2005), they are easy to write, easy to score (they can be scored automatically, in most cases), and can quickly provide feedback to instructors (e.g., how well students perform on a given item or which items are diagnostic of overall exam performance, as indicated by point-biserial correlations). This information can then be used to refine future tests, by eliminating overly difficult or easy questions, and/or eliminating items that are not diagnostic of exam performance. Furthermore, multiple-choice test questions can provide real-time feedback about performance in the classroom or during learning. For example, an instructor who gives multiple-choice clicker questions during class will be provided with immediate feedback on student performance. If the

majority of the class answers a question incorrectly, this can allow for remedial instruction on the concept, and this feedback is generally not afforded by short answer questions (particularly for large-enrollment classes). Secondly, a student taking a multiple-choice test provided by their textbook can immediately know if an answer is correct or incorrect, which can guide future study behavior. The ubiquity of multiple-choice tests in college classrooms is important when considering how educational techniques can use the forward testing effect in classrooms to their advantage. If recognition (e.g., multiple-choice questions) does not enhance new learning to the same degree as recall, which is relatively more effortful, then the forward testing effect may only play a role in educational contexts to the extent that classrooms or textbooks provide more difficult retrieval opportunities (e.g., recall) as opposed to easier retrieval opportunities (e.g., multiple-choice). Therefore, it is of utmost importance to determine the impact of interpolated multiple-choice testing using educational materials.

Differences between Recall and Recognition

Recall and recognition tests have long been thought to differentially rely on the processes of recollection and familiarity. Familiarity refers to a relatively automatic experience of knowing an item has been previously experienced and is generally thought to be less effortful relative to the process of recollection, which require effortful retrieval of previous study episodes (Jacoby, 1991). Some researchers have conceptualized recollection and familiarity as separate constructs (e.g., Jacoby, 1991; Jacoby & Dallas, 1981, Mandler, 2008), while others have conceptualized these as two endpoints of a continuum (Wais, Mickes, & Wixted, 2008; Wixted & Stretch, 2004). It is generally understood that successful recall usually relies heavily on recollection (but can benefit from familiarity), while recognition can be accomplished primarily by familiarity or some combination of the two processes, although one must exercise care in

equating tests with either process (Jacoby, 1991). In any case, recollection and familiarity differentially contribute to recall and recognition, with recollection contributing more to recall and familiarity contributing more to recognition. Therefore, the applied implications of the dual processes underlying recall and recognition are that recall tests (be they free-recall or cued-recall) are more "difficult" or effortful than recognition tests.[1]

With regard to the backward testing effect (i.e., retrieval of information increases the likelihood that the same information will be remembered later), Chan and McDermott (2007) examined how prior recall testing (relative to a no-test control) could impact later recognition performance, and found that final recognition (e.g., correct recognition of items that were previously presented) was largely insensitive to prior testing. In contrast, Roediger and Marsh (2005) examined what later benefit different types of *initial* testing (typically in a study-test-test/study-study-test design) could impart on performance on a final test. They found that initial multiple-choice testing resulted in a testing effect when the final test was cued-recall (i.e., more effortful) as opposed to when the final test was multiple-choice (similar to the findings of Chan & McDermott), but that lures from the initial test were likely to be repeated on the final test as errors. However, Butler and Roediger (2008) demonstrated that this detrimental effect of multiple-choice testing (e.g., repetition of lures on a final recall test) could be reduced with either immediate or delayed feedback.

In this vein, Bjork, Little, and Storm (2014) found that interpolated multiple-choice quizzes improved final exam performance in the classroom for both repeated questions and conceptually-related questions. This demonstrated that multiple-choice questions can be a useful learning tool in the classroom, although there was no comparison condition that used recall

[1] Recognition tests can encompass old-new recognition (as used in single-word lists), true/false questions, or alternative-forced choice (AFC) questions (i.e., multiple-choice).

(rather than recognition). Smith and Karpicke (2014), though, found that initial short-answer and multiple-choice tests in the laboratory led to similar-sized testing effects, unless short-answer questions were designed to elicit greater retrieval success (in which case these short answer questions resulted in larger testing effects than multiple-choice). In the classroom, McDermott, Agarwal, D'Antonio, Roediger, and McDaniel (2014) demonstrated a similar effect: initial short-answer and multiple-choice quizzes produced a similar benefit on final test performance, regardless of whether the initial and final test format (either short answer or multiple-choice) was the same. However, Carpenter and Delosh (2008) found that initial cued-recall tests (akin to short answer) resulted in *larger* testing effects than recognition, regardless of the final test format. The discrepancy between these studies may be due to the provision of corrective feedback after responding. When feedback is administered in a laboratory setting, participants are often given the correct answer as a correction. Studies that have found no difference in the testing effect between recall and recognition have typically administered feedback (e.g., Smith & Karpicke, 2014; McDermott et al., 2014), whereas those that have observed larger testing effects for recall over recognition have not (e.g., Carpenter & Delosh, 2008). In the case of the backward testing effect, feedback may equalize any differences between groups.

In contrast, for the forward testing effect, the difficulty of the initial or interpolated tests may play an important role in the size of the effect. For the most part, cued-recall tests should be more difficult than multiple-choice tests, resulting in poorer accuracy. Regardless of the presence of feedback, multiple-choice tests may lead to increased judgments of learning (JOLs), which could lead to the adoption of less effective study strategies on subsequent lists or text sections (Thiede & Dunlosky, 1994).

Therefore, while multiple-choice testing may strengthen *existing* memories as well as free-recall or cued-recall in the presence of feedback, this comparison has yet to be tested when it comes to potentiating *new* learning.

Proposed Mechanisms for the Forward Testing Effect

Prior Testing Improves Subsequent Study Strategies (A Metacognitive Account)

Of most importance for this dissertation is whether prior testing confers a metacognitive benefit that enhances new learning. Under this account, testing is proposed to enhance new learning by giving students accurate information about the nature and difficulty of the test(s). Learners can then use this information to encode the subsequent material more effectively (given their real or perceived performance on previous tests), which manifests as superior test performance during later interpolated tests and the criterial test. There are two critical assumptions of the metacognitive account. The first is that prior testing may alert students as to what sort of information may be tested and help alter their encoding strategy. For example, students who study only the big ideas for a test may subsequently alter their encoding strategy when they find that the actual test queried minute details contained in the text. Second, prior testing can provide metacognitive awareness of what is known and not known, which can drive subsequent re-encoding behavior (Davis & Chan, 2015; Dunlosky & Rawson, 2012; Kornell & Metcalfe, 2006). As an example, a student who does not test him- or her-self might be overconfident in their ability to recall material on a later test, leading to poorer test performance by way of less effective review strategies. A student who tests him- or her-self, in contrast, might understand how little of the material they understand, and adjust their study strategies accordingly to learn the material after the test. In both cases, what learners believe they know or do not know can influence their later encoding strategies.

As an illustration, Sahakyan, Delaney, and Kelley (2004) found that prior testing could enhance learning of a new set of materials in the context of a directed forgetting paradigm (see Figure 2). In the list-method directed forgetting paradigm, participants typically study two lists. After the first list they are instructed to remember or forget the previous list. Typically, participants told to forget the previous list demonstrate poorer recall of that list than those instructed to remember (the directed forgetting *cost*), while also remembering the second list better (the directed forgetting *benefit)*. Of particular interest is the benefit of directed forgetting, which is presumed to occur because proactive interference[2] (PI) is reduced by the forgetting instruction, However, In Sahakyan et al.'s experiments, participants were tested after the first list or not (after which they received the remember or forget instruction).

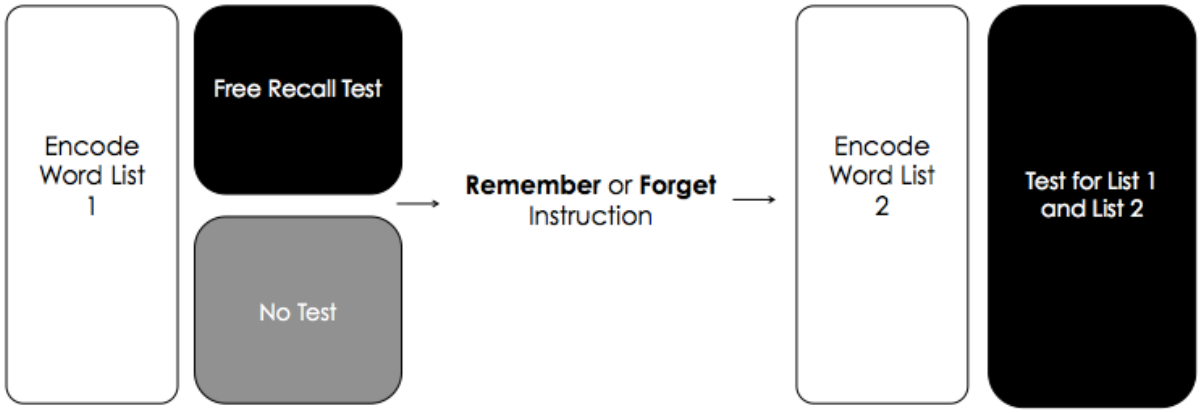

Figure 2. A graphical depiction of the method used by Sahakyan et al (2004).

The key finding is that prior testing eliminated the benefit (see Figure 3), but not the cost, of directed forgetting. The authors proposed that this occurred because the testing episode prompted a strategy change between the two encoding events, which then eliminated the benefit of the forget cue (which could have overridden the reduction of PI). This finding of enhanced

[2] The tendency for previously learned information to interfere with learning of new information.

learning of the second list was replicated by requiring participants to provide judgments of learning (JOLs) after the presentation of the first list. The authors argued that providing JOLs invoked metacognitive introspection, which led to better encoding (through a purported strategy change) of the second list. In a similar way, it is proposed that multiple-choice, relative to cued-recall, would be less likely to induce a strategy change, reducing its potential to enhance subsequent learning.

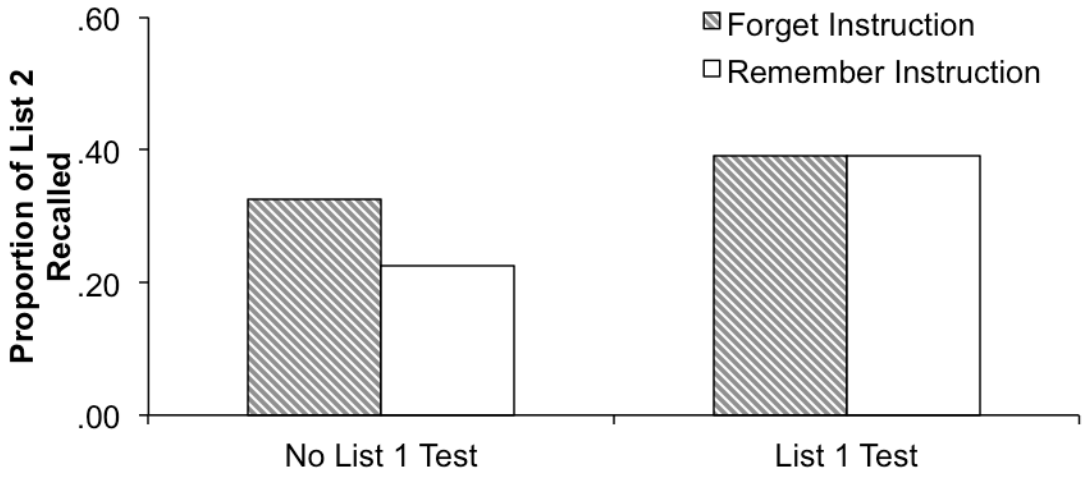

Figure 3. Results from Sahakyan et al (2004), Experiment 1.

Furthermore, much research has found that testing can reduce overconfidence in predicted performance. Students often overestimate their ability to remember information, and this over-confidence can lead to reduced study time (relative to students who are less confident) as well as less effective use of study habits (Dunlosky & Rawson, 2012). For example, students who study information may initially be overconfident in their ability to remember material. However, after a single test episode (Arkes, Christensen, Lai, & Blumer, 1987) this overconfidence is reduced relative to students who simply restudy the same information. This is particularly important as students use metacognitive judgments to decide what information to study (Kornell & Metcalfe, 2006). If students (incorrectly) assume that material is well-learned,

they may devote less study time to re-studying (or self-testing) in the future. As retrieval success on a multiple-choice test can be achieved primarily through familiarity, participants may not reduce their overconfidence in their performance to the same degree as participants who receive the relatively more difficult interpolated cued-recall tests. In the case of the forward testing effect, this means that any benefit of testing (compared to a no-test control) should be reduced when the interpolated tests are multiple-choice.

Yang, Potts, and Shanks (2017) had participants study several lists of different types of stimuli (Euskara-English word pairs, face-name pairs, and single word lists). As in the standard forward testing effect paradigm, all participants were tested after the final list (criterial test). Half of the participants were tested after the previous four lists while the other half completed math problems for the same amount of time after each list. The authors measured metacognitive beliefs in two ways. First, they directly measured aggregate judgments-of-learning (JOLs) after each list (Experiments 3 and 4), which reflected how well participants thought they would do on a later memory test for items from that list. Second, they allowed participants to proceed through each pair in each list at their own pace (Experiment 1 and 2). Therefore, participants could spend as long as they liked on learning each piece of information. This provided an indirect measure of metacognition, as participants should spend less time studying material if they are more confident in their learning, which may be the case when interpolated tests are not given (Koriat & Ackerman, 2010; Mazzoni, Cornoldi, & Marchitelli, 1990).

In fact, participants who had not been previously tested reported reduced JOLs across lists relative to participants who had been previously tested. Curiously, absolute correspondence between performance and predicted performance (the relation between predicted and actual performance, independent of under- or over-confidence) was the same for the tested and non-

tested groups for the final list, contrary to previous research (e.g., previously nontested participants did not demonstrate a larger difference between predicted and actual scores than previously tested participants, Dunlosky & Rawson, 2012). When encoding time was self-paced, non-tested participants spent less time across lists 1-5 encoding the word pairs. However, the encoding time of tested participants was relatively stable across lists. This reduction in encoding time for the non-tested participants suggests that, in the absence of interpolated tests, students may not effectively allocate their study time (either by paying attention while reading or engaging in more active strategies such as imagery generation or other mnemonics). These findings suggest that participants who were not previously tested might not have been aware of their poorer encoding strategies or the consequences thereof (e.g., poorer test performance)[3]. Essentially, tested learners in Yang et al's studies may have adopted better learning strategies relative to those learners who were not previously tested.

Szpunar, Jing, and Schacter (2014) asked for participants' judgments of learning after learning the entirety of a statistics video lecture (during which participants either took interpolated tests or did not). In contrast to Yang et al.'s (2017) findings (where JOLs were lower in the nontested than the tested group), the JOLs were roughly equivalent for participants who had received interpolated tests during the lecture and those who had not. This finding is interesting because previous research has found that repeated testing can reduce overconfidence, which should reduce JOLs (Dunlosky & Rawson, 2012). However, *actual* performance on the criterial test was drastically lower for participants who did not take interpolated tests on a final

[3] A corollary to this idea is the *motivational* hypothesis. The idea behind this hypothesis is that non-tested participants, without the reminder of periodic tests, are less motivated to perform well, and invest less effort in the task at hand. Whereas the metacognitive approach that I have proposed suggests that non-tested participants *may not know how* to effectively allocate study time or *may be oblivious* to their poor performance in the absence of retrieval practice, the motivational hypothesis suggests that non-tested participants may not be motivated to encode the material.

cumulative test. Therefore, correspondence between predicted and actual scores was higher for participants who took interpolated tests than those who did not. This is contrary to the findings of Yang et al, who found that correspondence between predicted and actual scores after the final *list* was equivalent between the tested and non-tested groups (although predicted and actual performance were lower for the non-tested group than the tested group). However, note that Yang et al. asked for aggregate JOLs after each list (e.g., four total), while Szpunar et al. asked for a single aggregate JOL after *all* materials had been studied. Providing judgments of learning is not necessarily a neutral event (Mitchum, Kelley, & Fox, 2016; Soderstrom, Clark, Halamish, & Bjork, 2015), and so the increase in JOLs for non-tested participants in Yang et al.'s study may reflect a reduction in the overconfidence with practice effect (Dunlosky & Rawson, 2012). This is not the only difference between the two studies, however. Yang et al. used Euskara-English word pairs, face-name pairs, and word lists, whereas Szpunar et al. used video-recorded sections of a statistics lecture as the to-be-learned material. Therefore, it could be the case that some materials more readily lend themselves to overconfidence than others. Regardless, the incongruity of results in the extant literature suggest that further study is needed to determine the effect of repeated testing (relative to no testing) opportunities on judgments of learning.

The primary aim of the current dissertation is to test the metacognitive account of the forward testing effect. Given the mixed results of previous research regarding metacognitive judgments and the forward testing effect, it is clear that further research is needed. In particular, the metacognitive account makes unique predictions regarding interpolated multiple-choice and interpolated cued-recall testing. Because retrieval success on a multiple-choice test is relatively easier to achieve than on a cued-recall test, accuracy on interpolated multiple-choice tests will likely be higher than that of cued-recall tests. Thus, I predicted that prior multiple-choice tests

should result in higher JOLs across learning opportunities, which in turn should lead to *poorer* performance on a cued-recall test for the final section (i.e., when the final section test is more difficult) relative to prior cued- recall tests. In contrast, prior cued-recall testing should result in decreased JOLs and a larger forward testing effect on a subsequent cued-recall test for the final section than prior multiple-choice testing. Lastly, a metacognitive account would predict little difference between prior multiple-choice testing and cued-recall testing when the final section test is multiple-choice, as learners in the latter condition will be over-prepared for the subsequent test while those in the former condition will be adequately prepared. One might predict that this over-preparedness of learners who had previously taken cued-recall tests would lead to better performance (relative to those who took multiple-choice tests) on a final multiple-choice test as well. However, given that multiple-choice tests are generally easier from a cognitive standpoint, they may be less sensitive to differences in prior encoding strategies (Chan & McDermott, 2007).

However, it is important to note that JOLs might not represent a neutral assessment of confidence in prior learning. There are some examples in the literature of JOLs influencing subsequent behavior outside of the effects of retrieval practice (see Rhodes, 2016 for a review). In some cases, providing evaluative judgments may influence *reactivity*, in which the prompt to reflect on future performance may results in participants' adjusting their encoding strategies on future learning opportunities. This may occur as a strategy change, in which a learner may switch from a less effective encoding strategy (e.g., rote repetition) to a more effective encoding strategy (e.g., generating imagery), or may reflect the suppression of negative behaviors (e.g., mind-wandering) that would otherwise occur throughout prolonged encoding. Typically, this reactivity manifests as superior performance in the control group, which ultimately masks a benefit of testing (but see Dougherty, Robey, & Buttacio, 2018, for an argument against this

proposition). Extant research using JOLs in the context of the forward testing effect have not found reactivity in response to JOLs (e.g., Yang, Potts, & Shanks, 2017). However, given the possibility that reactivity could occur, I tested learners in this paradigm either in the presence (Experiments 2 and 3) or absence (Experiment 1) of evaluative judgments.

Other Accounts of the Forward Testing Effect

A variety of other accounts have been proposed to explain the forward testing effect. Szpunar et al. (2008) proposed that interpolated testing reduces *proactive interference* (PI). In this case, interference can build up during studying information, such that information learned early in a study session may interfere with retrieval of material learned later. This effect becomes more pronounced as the length of a study set increases. Szpunar et al. observed that interpolated testing reduced this interference after each testing episode. Thus, after an interpolated test, students may approach each subsequent part of the text passage (and the subsequent test) as they do the first part (because they are relatively less encumbered by material from the previous section), which could enhance their learning of that material. Support for this proposition comes from the finding that testing not only enhances correct recall of new learning items on the final list but also *reduces* intrusions from prior lists. However, when comparing interpolated cued-recall and multiple-choice questions, it is unclear whether the reduction of PI account makes concrete predictions. Thus far, a specific mechanism for the manner in which tests reduce PI has not been specified, and such a specification would be necessary to make predictions in the context of the current studies.

A related idea is that interpolated testing isolates the context of each study list or text section (Bauml & Kliegl, 2013; Lehman, Smith, & Karpicke, 2014; Pastotter & Bauml, 2014; Sahakyan & Hendricks, 2012) and this contextual segregation helps constrain the retrieval set on

a later test. For example, in a lecture about lasers with no interpolated testing, all of the information learned in the lecture may fall under the context of the singular lecture. This large context may mean that retrieval of any of these events will be searched from this larger set, as in a chapter that contains no sub-headings. However, when testing is interpolated at several points during a lecture, these tests may serve to isolate different contexts during the lecture. For example, during retrieval, a student may think; "This was tested during the second section of lecture, which was over the laser mechanics." Now, the single lecture contains several contexts, including the events leading up to the development of the laser, the mechanics of the laser, and how lasers have been put to use in the public and private sphere. When questions are asked about each sub-section, particularly on the criterial test, students may be able to isolate the study context better, improving their performance.

It is also a possibility that interpolated testing produces the forward testing effect by helping students integrate new information with previously studied information. For example, Jing, Szpunar, and Schacter (2016) found that participants who had been tested during a lecture video tended to cluster their output (by recalling material together with other related material) more than participants who had not been tested. In other words, it appeared that participants had integrated information better after prior testing (relative to no prior testing). We (Chan, Manley, Davis, & Szpunar, 2018) have observed a similar pattern, with repeated test episodes associated with greater clustering of related words during recall.

Lastly, interpolated testing is associated with a number of positive learning behaviors, including reduced mind wandering, better note-taking, and increased attention. Szpunar and colleagues (2014) found that learners who received interpolated tests reported fewer bouts of mind-wandering (i.e., engaging in task-irrelevant thoughts), and Pastotter et al. (2011) found

neurological ERP evidence suggesting that learners pay attention during learning better following interpolated tests than following distractor tasks (but see Jing et al., 2016, for an exception where participants did not display a reduction in mind wandering after interpolated testing). In addition to reducing mind wandering during a lecture, Szpunar et al. (2013) found that participants who were tested on four sections of a lecture video took better notes than those who were only tested after the final section, suggesting that interpolated testing can enhance strategy use that leads to better encoding of subsequent material (as indexed by note-taking). Furthermore, Wilder, Flood, and Stromsnes (2001) found that giving unannounced extra-credit quizzes increased attendance by 10% relative to class periods[4] when students were aware that no quizzes would take place, and another study has found that frequent unannounced quizzes increased motivation to attend class and keep up with course material (Kouyoumdjian, 2004). Altogether, interpolated testing may reduce mind-wandering during lectures and text reading, as well as increase note-taking quality and attendance. These final points are critical, as interpolated testing may increase performance in the classroom without potentiating new learning through a cognitive or metacognitive mechanism at all, but by increasing "good student" behaviors that lead to better grades. The current experiments are not designed to test the predictions from these accounts. However, it is important to note that interpolated testing may influence behavior as well as memorial processes.

It is also important to note that the accounts of the forward testing effect (enhanced metacognition, reduction of PI, enhanced contextual segregation, and increasing good learning behaviors) are not mutually exclusive from one another. In fact, Chan, Meissner, and Davis

[4] This study followed a B-A-B design, in which each of four weeks of a course contained quizzes or not. Attendance was higher during four-week periods when students were expecting unannounced extra-credit quizzes as opposed to when they knew they would not be quizzed.

(2018) found varying levels of support for each of these accounts. Thus, it is entirely possible that each could contribute to the forward testing effect, albeit to varying degrees. However, the main focus of this dissertation is to examine how metacognitive knowledge acquired from different types of prior testing (relative to no testing) can guide subsequent encoding strategies. In particular, metacognitive judgments may be higher following prior multiple-choice testing, and this may lead participants to be less likely to adopt more effective encoding strategies for a section of text or a lecture that is tested with cued recall.

Effects of Prior Testing on a Cumulative Final Test

Another factor to consider when extrapolating this research to real-world learning environments is that one also needs to consider performance on a *final cumulative* test (heretofore referred to as a *follow-up test*). Therefore, this dissertation will examine how the forward testing effect persists on a final cumulative test for both previously tested (e.g., information that was tested during learning of the text sections) and previously non-tested information (e.g., information that was presented during learning, but not tested during the interpolated tests). For students and educators alike, performance on later formal assessments is likely of more interest than that of low-stakes interim quizzes, as high-stakes tests generally contribute more toward final course grades. If testing enhances new learning only on a criterial test, but fails to do so on a follow-up test, this may call into question the utility of interpolated tests as a learning tool to enhance new learning.

On this topic, results have been mixed. Some researchers have found that the forward testing effect persists in a final cumulative test (Jing et al., 2016, Szpunar et al., 2013), while others have found that it is diminished after a delay (Wissman & Rawson, 2015). It is important to note, however, that most research that has examined the forward testing effect has not

included a true no-test comparison condition (where participants aren't given the criterial test for the final section) so these final cumulative test data are almost always contaminated by the criterial test for the final section. Figure 4 displays graphically the difference between these two types of tests.

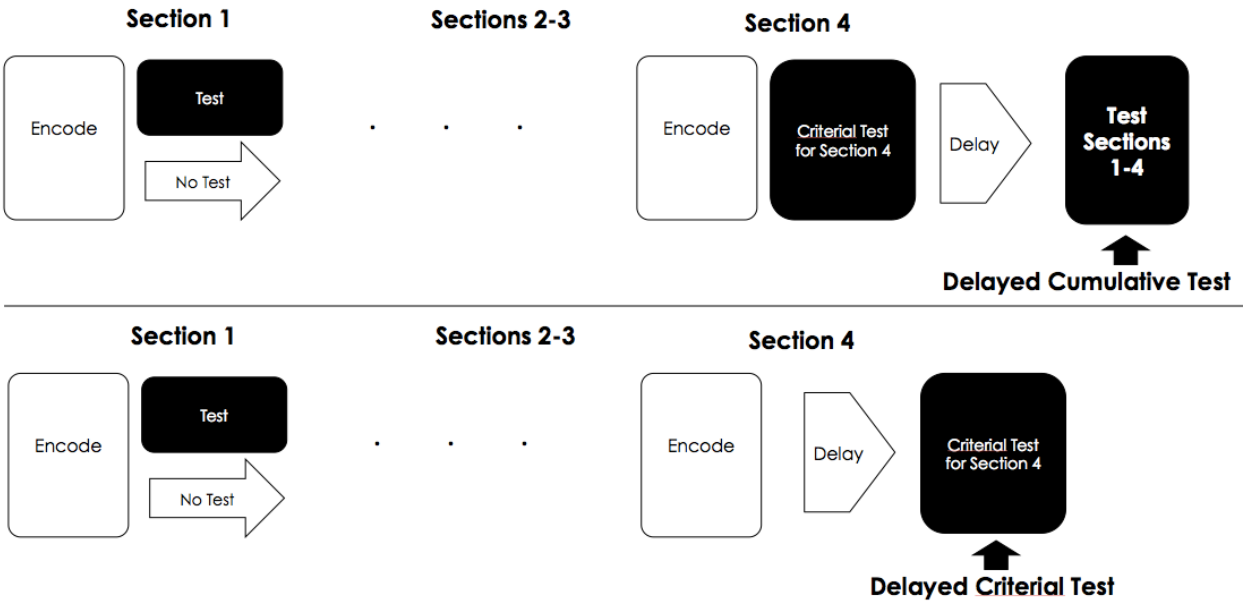

Figure 4. A graphical depiction of the difference between a delayed criterial test and a delayed cumulative test (which is contaminated by prior testing).

In a recent series of experiments, Chan, Manley, Davis, and Szpunar (2018) demonstrated that, in the absence of this contamination, the forward testing effect *did* persist across a delay, either 25 minutes prior to *testing* of the final list or 25 minutes prior to *encoding* of the final list. Note that this test was a criterial test, not a cumulative final test. In one case (Wissman & Rawson, 2015), performance was always contaminated by prior retrieval. However, a delayed criterial test (e.g., Chan, Manley, Davis, & Szpunar, 2018) is free from this contamination as long as there is no contamination from a prior criterial test.

However, in a real-world setting, exams are unlikely to be free from contamination by prior testing when interpolated tests are used in the classroom. If, as cognitive psychologists

suggest they do, students perform retrieval practice as a study tool and instructors implement in-class retrieval opportunities on a regular basis, the final exam will *always* be contaminated by prior retrieval *with feedback*. Therefore, it is important to understand how the forward testing effect influences final test performance when contamination is present. Wissman and Rawson (2015) proposed that any immediate benefit of testing should translate to a final test, based on the logic that greater initial test performance should persist (c.f., Rowland, 2014) to a later test. In their experiments, participants practiced retrieval after each section of text or after the entire text had been read. They predicted a main effect, such that enhanced performance after practice during section recall (relative to practicing the final section only) should carry over to the final test.

Contrary to their predictions, Wissman and Rawson (2015) found that the forward testing effect for the final passage was much smaller on a 20-minute delayed test than an immediate test. Essentially, both groups demonstrated backward testing effect of similar magnitudes, such that prior testing enhanced later performance relative to no prior testing. In contrast, Jing et al. (2016) found that the benefit of prior testing for the final lecture section persisted to a final cumulative test after a five-minute delay. Thus it remains unclear whether the early benefits of testing will persist on a follow-up test. No studies to date have examined delayed test performance using a criterial test that is not free recall; therefore, the present studies are novel in this respect (e.g. I used a cued-recall test as the final cumulative test). Furthermore, most studies examining the forward testing effect have not employed the use of feedback, which can play an important role in the size of the testing effect (Butler & Roediger, 2008). Therefore, in the present study, I predicted that test performance differences *on the criterial test*[5] would not translate to the follow-

[5] It is important to note that when discussing the forward testing effect persisting across a delay, that I am referring to the criterial test (e.g., Section 4 of the text). In this case, I predicted a forward testing effect immediately, such

up test, due to the presence of feedback provided on that criterial test (Kornell, Hays, and Bjork, 2008). Essentially, testing can enhance feedback learning regardless of the prior test condition, which may reduce any beneficial forward effect of testing.

A secondary question regarding final cumulative test performance is how students perform on items that were not queried on the interpolated tests when they encounter these items on a later cumulative test. Given the complexity of prose materials, it would be nearly impossible (and most certainly time-consuming) to test every single piece of encoded information. Therefore, a cumulative test may include repeated questions (e.g., questions that were previously tested) and non-tested questions (e.g., questions related to the tested material, but not tested on the interpolated tests). The literature is mixed as to the effects of selective (rather than comprehensive) retrieval practice. A common finding is that of retrieval-induced forgetting (RIFO[6].), in which selective retrieval of some items *impairs* later retrieval for the related non-tested items (see Anderson, 2003, for a review). However, retrieval-induced facilitation (RIFA) can also occur, in which prior selective retrieval practice *enhances* later retrieval for related material (Chan, 2009; Chan, McDermott, & Roediger, 2006). Given that the materials that generally demonstrate RIFA are more similar to the present materials (but see Saunders & MacLeod, 2002), I will discuss these latter findings in more detail here.

Chan and colleagues (2006, 2009) asked participants to read two text passages. For one of the passages, half of the facts presented in the passage were tested (e.g., Where do toucans

---

that criterial performance would be higher in the cued-recall condition than the multiple-choice or no-test condition, but that this difference would not persist to the follow-up test because *all* groups receive a test for section 4 and then receive feedback.

[6] In typical studies that have demonstrated RIFO, participants study lists of category-exemplar pairs (e.g., weather-tornado, weather-blizzard, fruit-banana, fruit-strawberry). They are then tested on some of the items from some of the categories (Rp+ items), but not on the other members of those same categories (Rp- items), or the other half of the categories (Nrp items). Thus, participants might practice "weather-to____", but not "weather-tsunami" or either of the "fruit" pairs. In this case, Rp- performance is typically worse than Nrp performance, a retrieval induced forgetting effect.

sleep? Answer: In tree holes), while the other related facts (e.g., What other species of bird are related to toucans? Answer: Woodpeckers) were not. No feedback was provided. In the other passage (which was unrelated to the first), none of the information was tested. Surprisingly, performance on a final test 24 hours later was higher for related non-tested items than for questions from the passage that were never practiced, suggesting that practicing some information from a prose passage can "spill over" to the learning of related material.

However, Chan (2009) demonstrated that there are important boundary conditions to this effect. Of most importance are two factors: 1.) Delay between initial and final test and 2.) How well the materials can be integrated together. Regarding the first factor, Chan only found RIFA when the delay between the initial test and final test was long (~24 hours), but not after a short delay of 20 minutes. Second, RIFA was only observed when the materials were easily integrated. For example, when sentences from the articles were presented in a random order (low integration condition), RIFO (e.g., forgetting) for the related non-tested questions occurred after a short 20-min delay, and there was no effect (i.e., neither forgetting nor facilitation) relative to unrelated non-tested questions after 24 hours. In contrast, when the sentences were structured in the originally written order (high integration condition) and participants were instructed to integrate them, there was no effect of selective retrieval on related items after the 20-min delay, but a RIFA (e.g., facilitation) effect was observed after 24 hours. Therefore, it appears that *both* a long delay and integration might be important factors that contribute to RIFA. In the present experiments, the cumulative final test occurred after only 15 minutes. However, given that the materials were presented in an order that is designed to enhance integration, I presumed that at least one factor that can facilitate RIFA was present in the current materials.

Regardless of the mechanism, one must presume that testing does not selectively enhance

learning of items that would be tested later. For example, if a participant is tested on four questions in Section 1, he/she should learn Section 2 better in general. For example, even though only four facts from Section 2 were *immediately* tested, all material (on average) should benefit from prior testing of Section 1, which should result in better performance on the follow-up test. Therefore, I predicted that interpolated testing would result in better performance for non-tested items relative to items from the no-test condition. At first glance, it seems as though it is difficult to disentangle this effect from RIFA[7]. This is certainly true, for all but the first set of materials. In fact, the only true test of RIFA is for material learned during the first passage of text, where there is no influence of prior testing.

Overview of Experiments

As mentioned previously, the goal of the present experiments was to evaluate how metacognitive knowledge contributes to the forward testing effect by employing several experimental procedures. First, I used multiple-choice tests to examine how variations in the difficulty of a test would affect the magnitude of the forward testing effect. I also required participants to provide overt judgments of learning (JOLs) after they had read each section of a passage in Experiments 2 and 3, which provides each participant's subjective level of mastery of the material. Participants studied four sections of a passage and were asked to judge immediately after reading each section, on a scale of 0-100, how likely they were to remember the material on an immediate cued-recall test, an immediate multiple-choice test, a later cued-recall test, and a later multiple-choice test. This allowed me to examine how prior testing (relative to no prior testing) influenced predictions of future performance. Second, I examined how long participants studied a section of a text passage before proceeding to the next section, which provides an

[7] Although it is unlikely for RIFA to emerge with a short delay like that used in the current experiments.

indirect measure of metacognition. If participants deem the material to be less difficult (which they may after receiving a prior multiple-choice test), they may spend less time studying the subsequent material. Yang et al. (2017) presented data on both reading times as well as aggregate JOLs in separate experiments. However, this dissertation is the first to collect both measures within a single experiment.

Combining both dependent measures in one experiment allows me to determine how metacognitive judgments influence self-regulation of encoding. For example, participants may increase their reading time across sections of text, and subsequently also increase their JOLs. In this case, they may be aware that they are engaging in more cognitive effort, which they expect to result in enhanced performance later (a positive correlation; more time spent reading will result in greater JOLs). However, the opposite could also be true. Participants may, based on the perceived ease of prior tests, assume that their performance will be better than it actually will be, and spend less time studying the material (a negative correlation, e.g., greater JOLs will result in less time spent reading). To my knowledge, no researchers have examined this correlation as it pertains to the forward testing effect. To examine this relation, I employed both cued-recall tests (which are more difficult), and multiple-choice tests (which are easier) to determine how prior testing influences both reading time behavior (an indirect measure of metacognition) and JOLs (a direct measure of metacognition).

A pilot experiment (N= 63) has provided some support for the first hypothesis: that multiple-choice testing results in a reduced forward testing effect relative to cued recall. In this pilot study, participants studied four sections of a prose passage about lasers. In all conditions, participants were given a 10-item cued-recall test after the fourth section (the criterial test). However, prior to the final section, participants either completed 10-item cued-recall tests for

each of the first three sections, completed multiple choice-tests, or read the passage in its entirety with instructions between each of the sections to advance to the next section. The results showed that cued-recall testing (M = .64) potentiated new learning relative to reading the full passage (M = .43), but multiple-choice testing did so to a lesser degree (M = .53). Note that this pilot experiment did not include conditions where participants took a multiple-choice test for the final section. Therefore, this study can only confirm the idea that multiple-choice tests may not potentiate new learning to the same degree as cued-recall when the final section test is also cued-recall.

As mentioned above, the criterial test in this dissertation was either cued-recall or multiple-choice. This allowed me to determine how a match or mismatch between previous tests and the test for the final section played a role in performance. Under a metacognitive account, prior multiple-choice testing should be a less potent instigator of the forward testing effect when the final test is a mismatch (e.g., when participants practice using multiple-choice for the first three sections and then take a cued-recall test for the final section) than when it is a match (e.g., when the tests for Sections 1-4 are all multiple-choice). Because recognition (e.g., multiple-choice) generally requires less cognitive effort than recall, participants who receive multiple-choice tests for the prior sections before a cued-recall test for the final section may be underprepared for the more difficult cued-recall test, due to the perceived ease of the previous tests. However, when the criterial test is also multiple-choice, participants may be adequately prepared for the types of question they will receive, which will result in equivalent criterial test performance regardless of prior test type. However, given the greater actual or perceived difficulty of retrieval under cued-recall conditions relative to recognition, I predicted that the prior tests to criterial section match may not play a role when the interpolated tests are cued-

recall (e.g., participants who receive prior cued-recall tests may be *over*-prepared for a multiple-choice test, and will perform just as well as those who were previously tested with multiple-choice[8]). It is also important to note that theories that account for the forward testing effect that propose proactive interference, integration, or contextual segregation as mechanisms are not mutually exclusive from a metacognitive account, and the present experiments were not designed to test between these different frameworks. Rather, the present experiments were designed to explore the metacognitive/strategy change account.

In Experiment 1, participants read through four sections of a text passage and either took a cued-recall test for Sections 1-3, took a multiple-choice test for Sections 1-3, or took no tests between the passages. Reading time was self-paced and the time spent reading each section was recorded. For the final section, all participants took either a cued-recall or a multiple-choice test. I also examined a final cumulative test performance 15 minutes later, and this test provides evidence for whether interpolated testing enhances learning on a later exam as well as whether interpolated testing influences learning of *non-tested* material. In Experiment 2, I required participants to provide overt judgments of learning after reading each section of the text passage in order to index their explicit metacognitive judgments. These JOLs were not collected in Experiment 1 because requiring JOLs during study may alter how learners approach the study task (namely, they may covertly retrieve information or evaluate their memory in order to provide the JOL, thus making it unclear whether later performance is due to factors that may not be present when JOLs are not required (Dougherty, Scheck, Nelson, & Narens, 2005; Mitchum et al, 2016; Sahakyan et al., 2004). Furthermore, providing judgments of learning could also

[8] If performance on the multiple-choice test for the final section is high, however, lack of a difference in the size of the forward testing effect based on the prior test may be a result of reduced sensitivity of the multiple-choice for the final section.

influence test expectancy (particularly in the no-test condition), which could eliminate the forward effect of testing. When frequently queried about future test performance, participants may begin to expect a test more often than when not given such questions about future performance.

Finally, in Experiment 3, I attempted to increase the potentiating effect of multiple-choice questions. To do so, I increased the difficulty of the multiple-choice questions by increasing the plausibility of the lures, which should reduce confidence, increase calibration, and result in enhanced learning of subsequent passages.

CHAPTER 2. EXPERIMENT 1

Method

Participants

All research for Experiments 1-3 was approved by Iowa State University's Institutional Review Board (see Appendix C). Two hundred and thirty participants were recruited from Iowa State University and received partial course credit for their participation. Data from 12 subjects were eliminated due to not finishing the experiment, eight were eliminated for indicating that English was not their native language, one was removed for having read the passage previously, and two online participants were eliminated because their reading times indicated that they had started the experiment but left the experiment for a duration of greater than 10 minutes. This yielded data from 207 participants overall, and approximately one-third of these were collected through the online SONA system at Iowa State University. Table 1 displays the number of participants per between-subjects condition as well as the exact number of online participants in each condition for Experiments 1-3. I determined minimum sample sizes based partially on the effect size (Hedge's g = .70) reported in a meta-analysis by Chan, Meissner, and Davis (2018). G*Power software indicated that, with an alpha value of .05, a power value of .80, and an expected effect size of .70, that 34 participants per between-subjects condition would be necessary to detect a single difference between two means.

Design

Experiment 1 used a 2 (Criterial Test Type: Cued Recall or Multiple Choice) x 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) between-subjects design. The primary dependent variable of interest was criterial test performance.

Table 1

Number of Participants in Each Between-Subjects Condition in Experiments 1-3

	Experiment		
Condition (PT-CT)	***1***	***2***	***3***
CR-CR	31 (5)	34 (10)	33 (10)
CR-MC	34 (9)	33 (10)	34 (10)
MC-CR	36 (12)	32 (10)	33 (10)
MC-MC	34 (8)	35 (10)	37 (10)
DMC-CR	--	--	39 (14)
DMC-MC	--	--	35 (11)
NT-CR	35 (11)	36 (9)	35 (11)
NT-MC	37 (10)	34 (10)	36 (12)

Note. Number of online subjects in each condition appears in parentheses. PT indicates prior test condition and CT indicates criterial test condition. CR indicates cued-recall, MC indicates (easy) multiple-choice, DMC indicates difficult multiple-choice, and NT indicates no test.

Materials

A 3,005-word prose passage (11.3 Flesch-Kinkaid grade level) was divided into approximately 750-word sections (see Appendix A). The passage discusses the development, use, and practical components of lasers. Each section was meant to be integrated with the others, such that learning one section should facilitate learning of the next section, and vice versa.

Note that while the passages do contain section headings in Appendix A, these headings were not presented to subjects as they read each section.

Eight open-ended test questions were designed for each section, and are presented in Appendix B. Half of the items served as test questions during the interpolated test, and half served as new related items on the later cumulative final test. For the cued recall tests, participants were prompted to provide a short response to the test prompt. In contrast, the multiple-choice tests presented the correct answer along with three foils (which were presented in a random order underneath the question), and participants were instructed to select the correct answer with a mouse click. Critically, the question stem was identical for both the multiple-choice and cued-recall questions. The only difference between the two conditions was that participants provided an open-ended response when the test was cued recall, but selected the response out of four options (plus an "I don't know" response option) when the test was multiple-choice. For example, one question queried, "What is a quantum?" In the short answer condition, participants were given a blank space to enter the correct answer (e.g., a fixed packet of energy). In the multiple-choice condition, participants were given the correct option along with three distractors (e.g., a unit of light, a unit of atoms, a number of electrons) For each section, half of the eight items were presented during the interpolated test, whereas the other half served as new questions on the final cued-recall test. During the interpolated tests, each test question was followed by the correct answer as feedback. For example, after the question about quanta, each participant would see the phrase "A quantum is a fixed packet of energy," before moving on to the next test question. Encoding of these feedback trials was self-paced, but the time spent encoding this feedback was not recorded.

Whether or not a given test question appeared in the interpolated tests was counterbalanced across subjects. The final test contained 32 cued-recall questions. Sixteen of these were repeated questions from the interpolated tests from sections 1-4. The test also included 16 new questions (four from each section) that were not queried in any of the interpolated tests. Thus, for participants in the multiple-choice and cued-recall groups, 16 of the items were repeated and 16 were new. For participants in the non-tested group, four items were repeated test questions (from the criterial Section 4 test) and 28 were new. There was no feedback provided after each test question in the cumulative test phase.

The distractor tasks before the final cumulative test were two working memory span tasks: alpha span and backward digit span. Collectively these tasks took approximately 15 minutes for subjects to complete. In the alpha span, subjects viewed a series of concrete nouns for 1.5 s each, and then were prompted to recall the items in alphabetical order (e.g., if a participant saw the words moose, gulf, and bear, they would be required to recall them in the order bear-gulf-moose). The number of items in a to-be-retrieved set ranged from two to five. Participants completed two trials for each set size and subjects saw the sets in ascending order of set size (e.g., participants first saw both 2-item set sizes, then both 3-item set sizes, and so on). During retrieval of a set, participants had five seconds per word to recall the items in the correct order. Thus, for a set size of four, a participant would have 20 seconds to respond, and would not be allowed to advance until that time had elapsed. The backward digit span task presented subjects with a series of digits (1-9) for 1.5 s each, and participants were required to retrieve them in the opposite order that they were presented (e.g., if a participant saw the digits 5, 9, 2 in that order, his/her task would be to recall the digits as 2, 9, 5). Similar to the alpha span, each

trial contained from two to seven digits, and participants completed two trials of each set size. Presentation of the trials proceeded from the smallest to the largest set size. During recall, participants had four seconds per digit in the task set.

Procedure

A graphical depiction of the procedure is presented in Figure 1. In all conditions, participants received instructions that they would be reading a series of text passages, and to try to learn them as if they were reading a textbook in one of their courses. Reading time was self-paced. Critically, participants were also told that after each passage, the computer would randomly determine how they would review the material they had just read: with a cued-recall test, a multiple-choice test, or no test. This instruction was important because it reduces the possibility that reduced test expectancy could cause poorer performance in the no-test condition (Weinstein, Gilmore, Szpunar, & McDermott, 2014). In actuality, participants performed the same review activity for the first three sections (depending on the assigned condition) and took either a multiple-choice test or a short-answer test for the fourth section. All participants were told that they would take a final cumulative test.

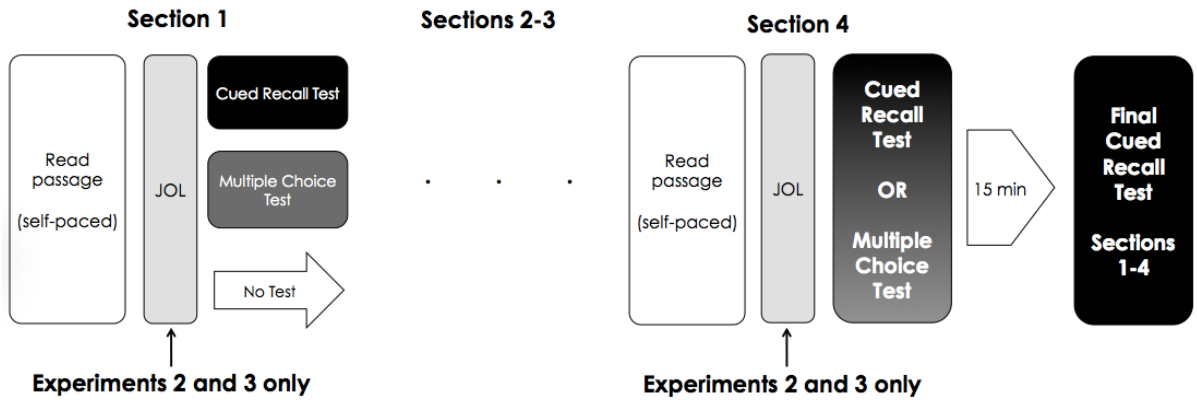

Figure 5. A graphical depiction of the method used in Experiments 1-3. Note that all interpolated tests for Sections 1-4 included feedback after each test question.

All materials were presented in Qualtrics survey software, and the reading of each passage was self-paced. After each of the first three sections, participants either took a four-question self-paced cued-recall test (e.g., "Who identified the two key components of the laser?"), took a four-question self-paced multiple-choice test that presents the correct answer (e.g., Albert Einstein) along with three foils (e.g., Isaac Newton, Gordon Gould, Stephen Hawking), or continued to the next passage. After encoding of the fourth and final section, all participants took either a four-question cued recall or multiple-choice test for that section. During the interpolated tests for Sections 1-4, all test questions were followed by self-paced correct feedback. Following completion of the criterial test for the final section, participants completed the alpha span and backward digit span tasks as distractors for 15 minutes before the final test[9]. After this retention interval, all participants received a final cumulative cued-recall test that included both repeated and new questions. This final test was self-paced, and questions were presented in a random order. Following the final test, all participants answered a short demographics questionnaire, which also queried what activity they had expected after reading the fourth passage in both an open-ended and forced-choice format.

Results

Interpolated and Criterial Test Performance

As can be seen in Figure 6, a 2 (Prior Review Type: Cued-Recall or Multiple-Choice) x 3 (Section 1-3) mixed ANOVA on the interpolated tests revealed that participants demonstrated poorer performance on cued-recall questions ($M = .24$) than multiple-choice questions ($M = .64$), $F(1,133) = 156.12$, $p < .001$, $\eta_p^2 = .54$. There was no effect of accuracy across sections, $F(2,$

[9] The tasks were timed to take approximately 15 minutes. However, participants may have differed in the amount of time spent reading instructions, which may result in the final retention interval varying between participants, although the exact retention interval for each participant was not recorded.

266) = 0.87 , p = .420, η_p^2 = .006, nor was there an interaction, $F(2, 266) = 0.73$, p = .485, η_p^2 = .005. This finding affirms that retrieval success was more difficult for cued-recall than multiple-choice questions.

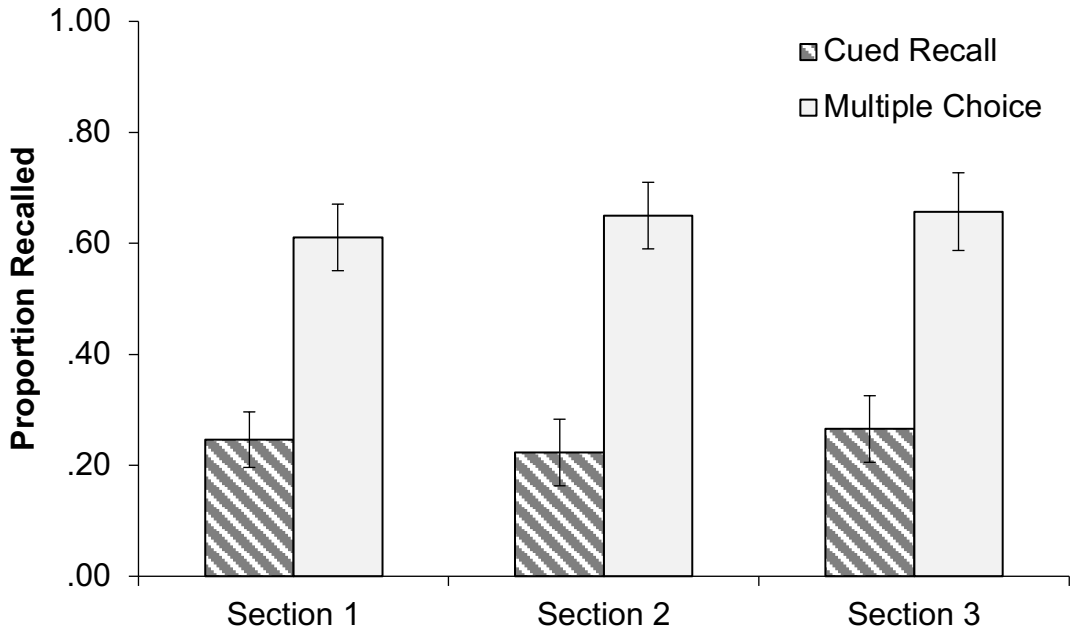

Figure 6. Interpolated test performance for Sections 1-3 as a function of interpolated test type in Experiment 1. Bars represent descriptive 95% confidence intervals.

Next, a 2 (Criterial Test Type: Cued Recall or Multiple Choice) x 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) between-subjects ANOVA was used to examine whether the forward testing effect was observed on the criterial test (see Figure 7). As in Sections 1-3, there was a main effect of Criterial Test Type, such that performance was higher overall when the criterial test was multiple-choice (M = .54) than when it was cued-recall (M = .41), $F(1, 201) = 10.42$, p = .001, η_p^2 = .049. There was also a main effect of Prior Review Type, $F(2, 201) = 6.47$, p = .002, η_p^2 = .060. Post-hoc tests revealed a moderately-sized forward testing effect when comparing the prior cued-recall (M = .56) to the no prior test condition (M = .39), $t(135) = 3.42$, p = .001, d = 0.59. When comparing the prior multiple-choice (M = .48) to the no

prior test condition, there was a marginal forward testing effect $t(140) = 1.95$, $p = .053$, $d = 0.33$, although this was nearly half the size of the effect for prior cued-recall. The difference between prior cued-recall and prior multiple-choice was also only marginally significant, $t(133) = 1.69$, $p = .094$, $d = 0.29$.

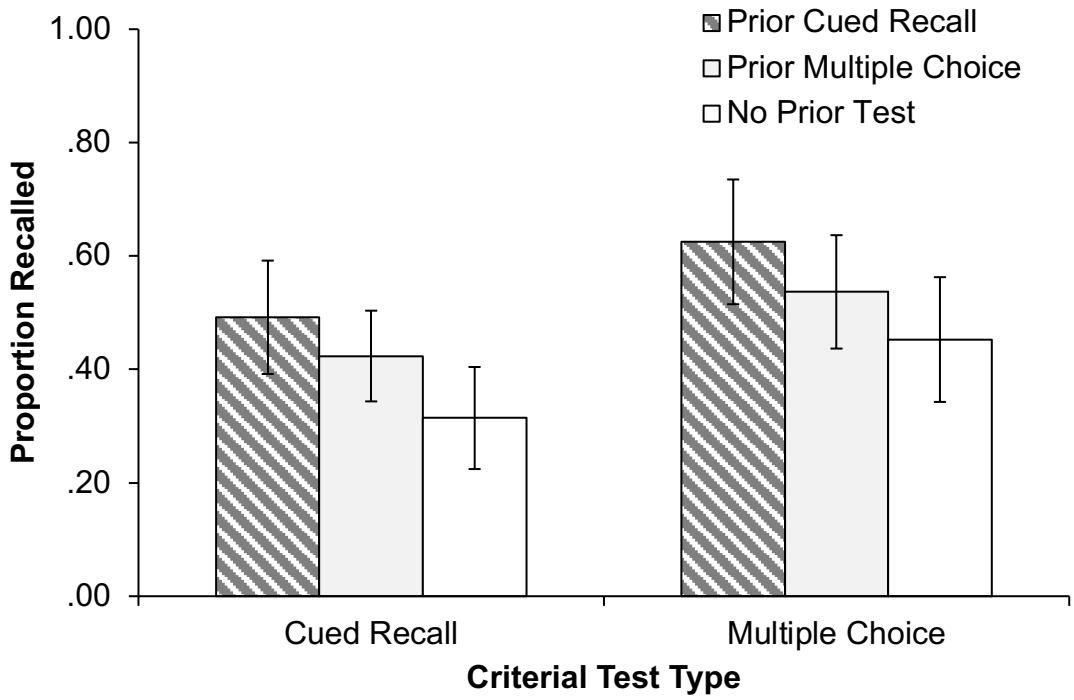

Figure 7. Criterial test performance for Section 4 as a function of prior interpolated test type in Experiment 1. Bars represent descriptive 95% confidence intervals.

Section 1-4 Reading Times

A 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) x 4 (Section: 1-4) mixed ANOVA was used to examine the reading times for each of the four passages (see Figure 8). There was a main effect of Prior Review Type, $F(1, 204) = 6.30$, $p = .002$, $\eta_p^2 = .058$. There was no effect of Section, $F(3, 612) = 1.78$, $p = .149$, $\eta_p^2 = .009$, nor was there an interaction, $F(6, 612) = 1.66$, $p = .128$, $\eta_p^2 = .016$. Post-hoc tests determined that on average, participants spent more time reading in the prior cued-recall condition ($M = 224$s*)* than in the no prior test

condition (M = 172s), $t(135) = 3.00$, $p = .003$, $d = 0.52$, and more time reading in the prior multiple-choice condition (M = 247s) than the prior no-test condition $t(140) = 3.21$, $p = .002$, $d = 0.54$. There was no difference in average reading time between the prior cued-recall and prior multiple-choice conditions, $t(133) = 0.96$, $p = .337$, $d = 0.17$. Based on a priori predictions, I also tested the linear trend across Sections 1-4 separately for each between-subjects condition by conducting linear contrasts. There was no linear trend for the prior cued-recall condition, $F(1, 64) = 0.67$, $p = .415$, $\eta_p^2 = .010$, or for the prior multiple-choice condition, $F(1, 69) = 0.29$, $p = .592$, $\eta_p^2 = .004$, indicating that reading times remained stable across sections for both tested

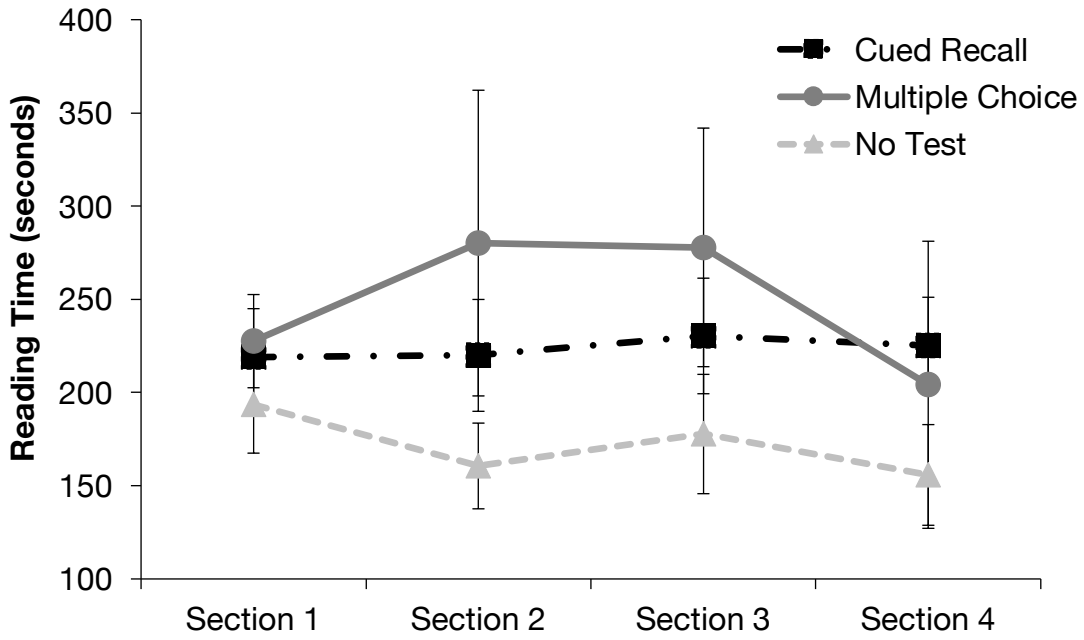

Figure 8. Reading time as function of prior interpolated test condition for Sections 1-4 in Experiment 1. Bars represent 95% confidence intervals.

conditions. However, reading times declined in a linear fashion across sections in the no prior test condition, $F(1, 71) = 5.39$, $p = .023$, $\eta_p^2 = .071$. Post-hoc tests revealed that reading times declined from Section 1 (M = 193s) to Section 2 (M = 161s), $t(71) = 3.69$, $p < .001$, $d = 0.44$, increased slightly but not significantly from Section 2 to Section 3 (M = 178s), $t(71) = 1.41$, p =

.163, $d = 0.17$, and decreased marginally from Section 3 to Section 4 ($M = 156$s), $t(71) = 1.92$, $p = .057$, $d = 0.23$. Therefore, it appears that the negative linear trend is primarily due to a decrease in reading times from the first to the second section of the passage. Overall, the reading time results suggest that learners in the no-prior test condition begin to spend less time encoding across passages after Section 1, but that tested learners remain relatively stable in their reading times.

I also examined the correlation between reading times for Section 4 and accuracy on the criterial test. When the criterial test was cued-recall, the correlation was not significant, $r = .03$, $p = .805$, nor was it significant when the criterial test was multiple-choice, $r = .16$, $p = .105$.

Cumulative Test Performance

First, a one-way ANOVA was performed on overall test performance (collapsing across Sections and old and new items) between the three groups that received cued-recall, multiple-choice, or no-test during the non-criterial practice phase (see Table 2). Overall test performance was composed of old and new items from Sections 1-4. Recall that for the previously tested conditions, 16 of 32 items on this test were old, and 16 were new. In the no prior test condition, 4 items were old (i.e., the four items tested on the criterial test) and 28 items were new (e.g., all of the items from Sections 1-3 and four items from Section 4). This yielded a main effect of prior test condition $F(2, 207) = 22.45$, $p < .001$, $\eta_p^2 = .178$. Post-hoc tests revealed that participants who previously received cued-recall tests ($M = .40$) for Sections 1-3 outperformed those who previously took no tests ($M = .27$) for Sections 1-3, $t(138) = 6.11$, $p < .001$, $d = 1.04$. A similar testing effect was observed for the participants who previously took multiple-choice tests ($M = .40$) for Sections 1-3, $t(142) = 5.80$, $p < .001$, $d = 0.97$. There was no difference in cumulative test performance between the two previously tested groups $t(134) = 0.10$, $p = .921$, $d = 0.02$.

These analyses confirm a backward testing effect between-subjects, such that individuals who received interpolated tests perform better on a final cumulative test than those who did not receive interpolated tests.

Table 2

Cumulative Test Performance as a Function of Prior Interpolated Test Condition for Old and New Items from Sections 1-3

	Section		
	1	***2***	***3***
Cued Recall			
Old	.47 *(.03)*	.55 *(.03)*	.47 *(.04)*
New	.18 *(.03)*	.25 *(.03)*	.20 *(.03)*
Multiple Choice			
Old	.50 *(.03)*	.49 *(.03)*	.46 *(.03)*
New	.23 *(.02)*	.23 *(.03)*	.15 *(.02)*
No Test			
Old	--	--	--
New	.18 *(.02)*	.15 *(.03)*	.20 *(.02)*

Note. Numbers in parentheses represent standard errors of the mean.

A second analysis was performed on the items that were previously tested or not tested in Sections 1-3 for each tested condition. For the two prior tested conditions, half of the items on the cumulative test were new, and half were old. In the no prior test condition, all of the items were new (see Table 2), so these data were not included in the analysis. A 2 (Item Type: Previously Tested or New) x 2 (Prior Review Type: Cued Recall or Multiple Choice) mixed ANOVA yielded a main effect of item type, such that old items (M = .49) were recalled more often than new items (M = .21), $F(1, 134) = 303.58$, $p < .001$, $\eta_p^2 = .694$. This represents a within-subjects testing effect.

There was no main effect of Prior Review Type, $F(1, 134) = 0.06$, $p = .805$, $\eta_p^2 < .001$, nor was there an interaction, $F(1, 134) = 0.02$, $p = .892$, $\eta_p^2 < .001$.

I also examined test performance for new items only in a 3 (Section: 1-3) x 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) mixed ANOVA. This yielded a main effect of Section, $F(2, 414) = 3.14$, $p = .044$, $\eta_p^2 = .015$, no main effect of Prior Test Condition, $F(2, 207) = 0.49$, $p = .616$, $\eta_p^2 = .005$, and a significant interaction, $F(4, 414) = 2.75$, $p = .028$, $\eta_p^2 = .026$. To examine the main effect of Section, I compared performance from Section 1 ($M = .21$) to Section 2 ($M = .22$), which did not differ, $t(209) = 0.49$, $p = .626$, $d = 0.03$. There was a significant but small drop in performance from Section 2 to Section 3 ($M = .18$), $t(209) = 2.32$, $p = .022$, $d = 0.16$. There was also a small marginal drop in performance between Section 1 and Section 3, $t(209) = 1.81$, $p = .072$, $d = .12$. These comparisons demonstrate that the items from Section 3 may have been slightly more difficult than the other sections. To decompose the interaction, I conducted separate one-way within-subjects ANOVAs for each prior test condition. The effect of Section was not significant in the prior cued-recall condition, $F(2, 130) = 2.08$, $p = .130$, $\eta_p^2 = .031$, nor was it significant in the no prior test condition $F(2, 146) = 0.86$, $p = .427$, $\eta_p^2 = .012$. There was a significant effect in the prior multiple-choice condition, $F(2, 138) = 5.22$, $p = .007$, $\eta_p^2 = .070$. Post-hoc *t*-tests revealed that there was no difference in performance between Sections 1 ($M = .23$) and 2 ($M = .23$), $t(69) = 0.00$, $p = 1.000$, $d = 0.00$, but that performance in Section 3 ($M = .15$) was lower than that of Section 1, $t(69) = 3.03$, $p = .003$, $d = .36$, and Section 2, $t(69) = 2.67$, $p = .010$, $d = .32$. This signifies that those in the prior multiple-choice condition demonstrated a sharper drop in performance for Section 3 compared to the other prior review conditions (cued-recall and no-test).

Finally, I analyzed cumulative test performance for items from the criterial test in a 2 (Item Type: Previously Tested or New) x 3 (Prior Review Type: Cued Recall, Multiple Choice, or No-Test) x 2 (Criterial Test Type: Cued Recall or Multiple Choice) mixed ANOVA. As can be seen in in Table 3, there was a within-subjects testing effect, with old items (M = .77) remembered more often than new items (M = .35), $F(1, 207) = 388.58$, $p < .001$, $\eta_p^2 = .652$. There was no effect of Prior Test Condition, $F(2, 207) = 0.49$, $p = .614$, $\eta_p^2 = .005$. There was a significant interaction between Item Type and Criterial Test Type, $F(1, 204) = 4.82$, $p = .012$, $\eta_p^2 = .031$. The two-way interactions between Prior Review Type and Criterial Test Type, $F(2, 207) = 2.30$, $p = .103$, $\eta_p^2 = .022$, Item Type and Prior Review Type, $F(2, 204) = 1.65$, $p = .113$, $\eta_p^2 = .021$, and the three-way interaction, $F(2, 204) = 0.17$, $p = .841$, $\eta_p^2 = .002$ were not significant.

Table 3

Cumulative Test Performance as a Function of Prior Interpolated Recall Condition and Criterial Test Type for Old and New Items from Section 4.

	Criterial Test Type	
Prior Review Type	***Cued Recall***	***Multiple Choice***
Cued Recall		
Old	.76 *(.04)*	.84 *(.03)*
New	.33 *(.05)*	.32 *(.04)*
Multiple Choice		
Old	.72 *(.05)*	.79 *(.03)*
New	.43 *(.05)*	.35 *(.04)*
No Test		
Old	.74 *(.04)*	.78 *(.03)*
New	.36 *(.04)*	.31 *(.04)*

Note. Numbers in parentheses represent standard errors of the mean.

Post-hoc tests comparing the two levels of the Criterial Test Type variable separately for old and new items were conducted to decompose the significant two-way interaction. These tests revealed that performance was better for old items when the criterial test was multiple-choice (M = .80) compared to cued-recall (M = .74), $t(208) = 2.13$, $p = .034$, $d = 0.30$, but that there was no difference for new items (Ms = .33 and .38, respectively), $t(208) = 1.33$, $p = .185$, $d = 0.18$. This signifies that participants might have better learned the feedback from the criterial test when they were tested with multiple-choice than cued-recall.

Test Expectancy

Table 4

Section 4 Test Expectations as a Percentage of Each Prior Test Condition

		Expectation	
Prior Test Condition	***Cued Recall***	***Multiple Choice***	***No Test***
Cued Recall	65%	15%	20%
Multiple Choice	27%	59%	14%
No Test	25%	48%	24%

After the cumulative test, participants were asked to indicate which activity they had expected after the reading of the final section of the passage. Participants were first asked what they had expected as an open-ended question. They were then were given a forced choice question, which included the options "Short Answer Test," "Multiple Choice Test," or "No Test." As responses to the open-ended question indicated that participants did not understand the instructions (e.g., participants reported "expecting" the distractor tasks that occurred after Section 4), only data from the forced-choice question are presented here. These data are presented in Table 4. These data are exploratory, so no formal analysis was used to analyze

these data. However, one can glean from examining Table 4 that individuals who previously received cued-recall tests were more likely to report expecting cued-recall, and individuals who were previously tested with multiple-choice were more likely to report expecting a multiple-choice test. Strangely, those in the no-test condition were more likely to report expecting a multiple-choice test as well. However, these data were collected retrospectively, so interpreting this result is somewhat difficult. Regardless, among the tested participants, it seems that learners expected the task that they had previously done.

Discussion

There are two key findings that bolster the metacognitive account of the forward effect of testing in Experiment 1. First, reading times remained consistent across text passage sections in the two tested conditions, but dropped after the first section in the non-tested condition. This suggests that prior testing may encourage learners to continue to pay attention to each passage in order to effectively encode the information. Second, and most importantly, prior cued recall testing moderately enhanced new learning *regardless of the criterial test type*. While prior multiple-choice testing did enhance learning of the final section to a small degree, the effect size was nearly half that of cued-recall. This experiment is the first to demonstrate that more difficult tests (e.g., cued-recall) may have greater effects on *future* learning. This is consistent with work in the backward-testing effect, where prior difficult retrieval results in better retrieval later on (Carpenter & Delosh, 2006). Taken together, it appears that both test types can enhance positive study behaviors, but that a more difficult test (i.e., cued-recall) produces a greater forward testing effect.

However, the results from the final test suggest that the forward testing effect may be short lived. While there were large between-subjects and within-subjects backward testing

effects, the initial forward testing effect from the criterial test was absent on the cumulative test after a 15-minute delay. It is important to note that the criterial test provided feedback, and as all conditions were tested, this could have contributed to the lack of a persistent forward testing effect for the old items. Thus, even individuals in the previously non-tested condition received feedback for old items, and could have learned the feedback to the same degree as those who were previously tested. However, the lack of a forward testing effect for new items (e.g., comparing criterial test recall for new items from either tested condition to the non-tested condition) suggests that the forward testing effect may not persist across a delay for these types of materials. In one case, Wissman and Rawson (2015, using prose passages) found that the forward testing effect was reduced (but not eliminated) after a 20-minute delay, although this is inconsistent with recent work which has demonstrated that the forward testing effect remains at a similar magnitude after a 25-minute retention interval (Chan, Manley, Davis, and Szpunar, 2018, using word lists) It may be the case that different types of materials show differential effects of retention intervals, or it may be the case that the forward testing effect is fragile over time.

This experiment was the first to empirically test the effectiveness of interpolated multiple-choice tests against cued-recall tests in regards to the forward testing effect. However, to determine whether the reduced forward effect of testing with multiple-choice questions is indeed related to metacognitive overconfidence, it is necessary to require participants to provide explicit estimates of performance after reading each passage. Thus, in Experiment 2, I required participants to provide aggregate judgments of learning (JOLs).

CHAPTER 3. EXPERIMENT 2

Method

Participants, Design, Materials, and Procedure

Two hundred and twenty-two participants were recruited from Iowa State University and received partial course credit for their participation. Five participants were eliminated from the analysis because they did not finish the experiment, and 11 were eliminated for indicating that English was not their native language. This yielded 204 participants for the final analysis. Minimum sample sizes were based on the same power analysis conducted in Experiment 1, and the number of participants in each between-subjects condition (as well as the number of online subjects) is displayed in Table 1.

Experiment 2 uses a 2 (Criterial Test Type: Cued Recall or Multiple Choice) x 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) between-subjects design, and the primary dependent variable of interest was criterial test performance.

The procedure (see Figure 1) was nearly identical to that of Experiment 1, with one notable exception. In Experiment 1, participants proceeded immediately from the section reading to the section test (or simply proceeded to the next section in the No-Test condition). In Experiment 2, however, participants were prompted to provide four aggregate JOLs (i.e., estimate their future performance), one for an immediate cued-recall test, one for an immediate multiple-choice test, one for a final cumulative cued-recall test, and one for a final cumulative multiple-choice test. Each question asked participants to indicate their JOL by using a sliding scale from 1-100, and these judgments were provided immediately after they had finished reading each section of the text. Specifically, participants were told, “Please now indicate how well you will perform on a future test for the material you just read: 1.) How much, on a scale of

1-100, do you think you will remember on a short-answer test *right now?* 2.) How much, on a scale of 1-100, do you think you will remember on a multiple-choice test *right now?* 3.) How much, on a scale of 1-100 do you think you will remember on a short-answer test in 25 minutes? 4.) How much, on a scale of 1-100 do you think you will remember on a multiple-choice test in 25 minutes? The four JOL questions appeared one at a time on the screen, and the order in which they appeared was randomized to reduce the likelihood that participants would anchor their JOLs on any one judgment.

Results

Interpolated and Criterial Test Performance

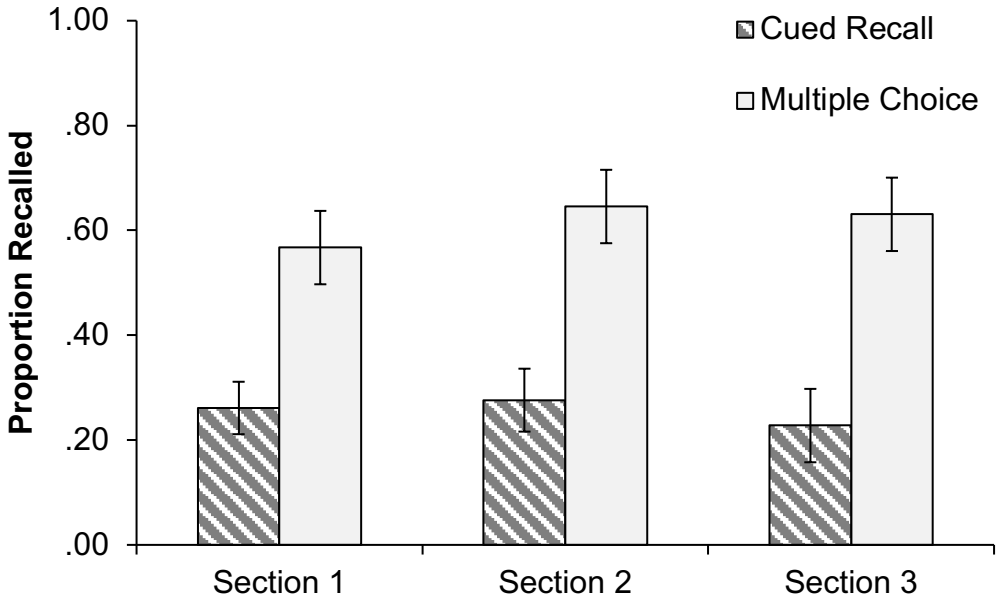

Figure 9. Interpolated test performance for Sections 1-3 as a function of interpolated test type in Experiment 2. Bars represent descriptive 95% confidence intervals.

For the interpolated tests, a 2 (Prior Review Type: Cued-Recall or Multiple-Choice) x 3 (Section 1-3) mixed ANOVA revealed that performance was higher on the interpolated multiple-choice tests (M = .61) than for the interpolated cued-recall tests (M = .25), $F(1, 132) = 115.00$, p

< .001, η_p^2= .836 (see Figure 9). There was no main effect of Section, $F(2, 264) = 1.84$, p = .160, η_p^2 = .014, nor was there an interaction $F(2, 264) = 1.97$, p = .177, η_p^2 = .015. This finding replicates the results from Experiment 1 and demonstrates that the interpolated cued-recall tests were more difficult than the interpolated multiple-choice tests.

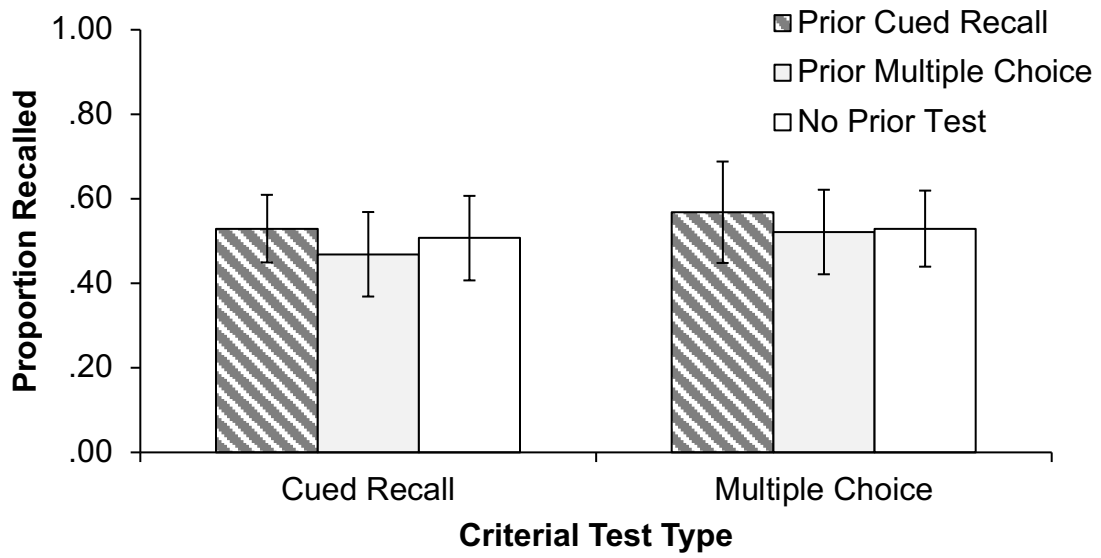

Figure 10. Criterial test performance for Section 4 as a function of prior interpolated test type in Experiment 2. Bars represent descriptive 95% confidence intervals.

A separate 2 (Criterial Test Type: Cued Recall or Multiple Choice) x 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) between-subjects ANOVA was used to examine the forward testing effect on the criterial test (see Figure 10). Surprisingly, there was no effect of Prior Review Type, $F(2, 198) = 0.62$, p = .545, η_p^2 = .006, Criterial Test Type, $F(1, 198) = 0.92$, p = .339, η_p^2 = .005, nor was there an interaction, $F(2, 198) = 0.05$, p = .952, η_p^2 < .001. In other words, there was no forward effect of testing observed in Experiment 2, nor was there an effect of multiple-choice vs. cued-recall difficulty on the criterial test. It is a possibility that requiring participants to provide JOLs may have increased test expectancy or induced

metacognitive introspection that induced a strategy change. I will return to these two null effects in the General Discussion.

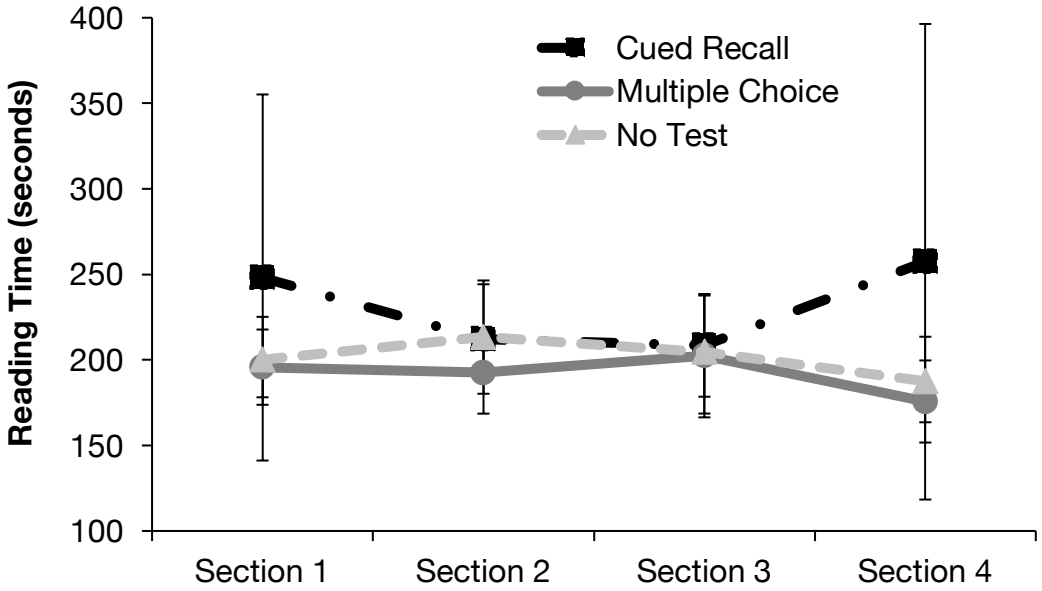

Figure 11. Reading time as function of prior test condition for Sections 1-4 in Experiment 2. Bars represent descriptive 95% confidence intervals.

Section 1-4 Reading Times

In order to analyze how reading times (see Figure 11) were affected by prior review condition, I ran a 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) x 4 (Section: 1-4) mixed ANOVA. There was no effect of Prior Review Type, $F(1, 201) = 1.45$, $p = .237$, $\eta_p^2 = .014$, no effect of Section, $F(3, 603) = .086$, $p = .968$, $\eta_p^2 < .001$, nor was there an interaction, $F(6, 603) = .60$ $p = .734$, $\eta_p^2 = .006$. Despite these null effects, I also conducted linear trend tests separately for each Prior Review Type because of my a priori prediction that the no-test condition would exhibit a negative linear trend in reading times across sections. In fact, no linear trend was observed in the prior cued-recall, $F(1, 66) = 0.01$, $p = .928$, $\eta_p^2 < .001$, prior multiple-choice, $F(1, 66) = 1.73$, $p = .192$, $\eta_p^2 = .026$, or in the no prior test conditions, $F(1, 69) = 1.71$, $p = .196$, $\eta_p^2 = .024$.

Thus, unlike in Experiment 1, testing (or no testing) had no impact on reading times, which remained stable across sections of the text for each condition.

I also examined the correlation between reading times and performance on Section 4 for each criterial test type. For the criterial cued-recall test, the relation between reading time and performance was not significant, $r = -.15$, $p = .221$, although it was marginal for the criterial multiple-choice test, $r = .22$, $p = .079$.

Judgments of Learning

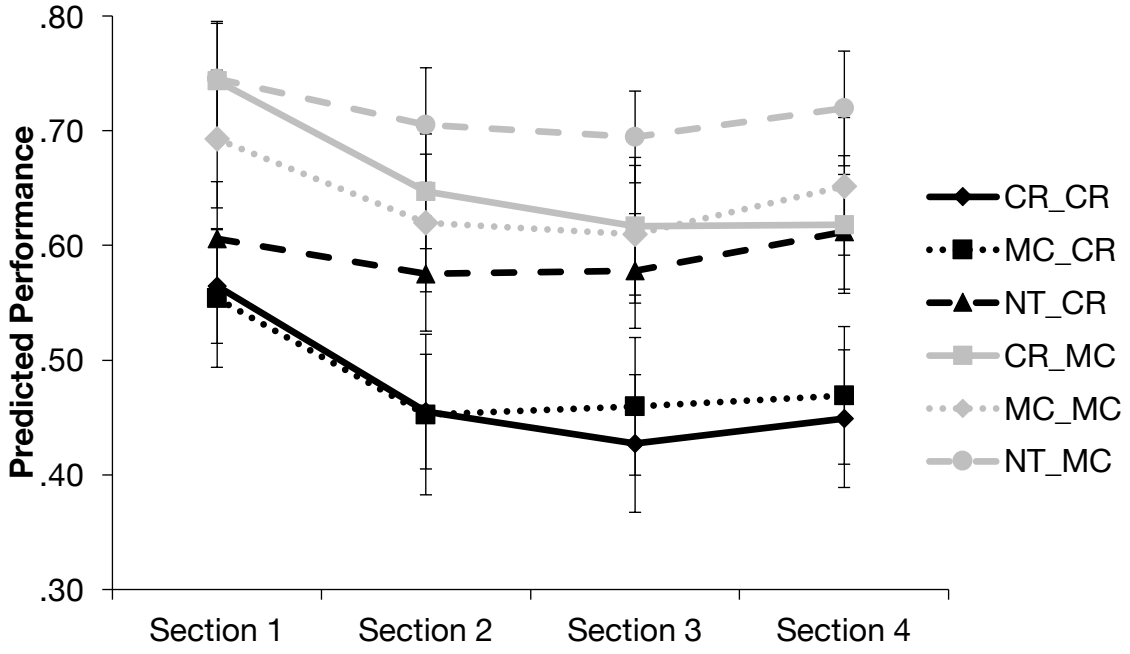

Figure 12. JOL's provided by subjects for Sections 1-4 in Experiment 2 as a function of prior testing and type of test prediction. The first denotation in each condition refers to the prior test condition (CR = cued-recall, MC = multiple-choice, and NT = no test), and the second denotation refers to the type of test prediction. Gray lines indicate predictions made regarding future multiple-choice performance, while black lines indicate predictions made regarding future cued-recall performance.

The JOLs for each section were analyzed using a 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) x 4 (Section: 1-4) x 2 (JOL Test Type: Cued-Recall or Multiple-

Choice) mixed ANOVA. Recall that before each test, participants provided four aggregate JOL judgments: one for the initial multiple-choice test, one for the initial cued-recall test, and one for the later cumulative cued-recall test. For the purposes of this analysis, only the predictions for the immediate tests have been used. As can be seen in Figure 12, the ANOVA yielded a main effect of Section, $F(3, 603) = 27.28$, $p < .001$, $\eta_p^2 = .119$, a main effect of JOL Test Type, $F(1, 201) = 320.21$, $p < .001$, $\eta_p^2 = .614$, and a main effect of Prior Review Type, $F(2, 201) = 6.11$, $p < .001$, $\eta_p^2 = .057$. The two-way interaction between Section and Prior Review Type was significant, $F(6, 603) = 3.76$, $p = .001$, $\eta_p^2 = .036$, as was the interaction between JOL Type and Prior Review Type, $F(2, 201) = 3.97$, $p = .020$, $\eta_p^2 = .038$. The two-way interaction between Section and JOL Type was not significant, $F(3, 603) = 0.39$, $p = .761$, $\eta_p^2 = .002$, nor was the three-way interaction, $F(6, 603) = 1.42$, $p = .204$, $\eta_p^2 = .014$.

In order to understand the main effect of Section, I first conducted a linear contrast for the Section variable collapsed across the other variables. This contrast was significant, $F(1, 201) = 28.02$, $p < .001$, $\eta_p^2 = .12$. Inspecting Figure 12, one can see that this linear trend is in the negative direction, with JOLs decreasing from Sections 1-4[10]. Figure 12 also shows that the main effect of JOL Test Type is due to higher estimates of performance for multiple-choice tests ($M = .67$, gray lines) than cued-recall ($M = .52$, black lines), suggesting that participants were indeed aware of the increased difficulty of the cued-recall tests and adjusted their estimates of performance accordingly. Last, JOLs were higher for those who had previously not taken any interpolated tests ($M = .65$) relative to those who had taken cued-recall tests ($M = .57$), $t(135) = 3.22$, $p = .002$, $d = 0.55$, and relative to those who had taken multiple-choice tests ($M = .56$), $t(135) = 2.90$, $p = .004$, $d = 0.50$. There was no difference in JOL's between the two tested

[10] Note, however, that each section contained different materials. Thus, the main effect of Section on its own is less informative than examining the differences between the levels of the Prior Test Type and JOL Test Type variables.

conditions, $t(132) = 0.05$, $p = .958$, $d = 0.01$. This suggests that learners who were tested had reduced confidence in future performance while learners who were not tested were more confident in their future performance predictions.

To decompose the interaction between Section and Prior Review Type, I conducted linear trend tests separately for each condition across sections. There was a linear trend for the prior cued-recall condition, $F(1, 66) = 19.58$, $p < .001$, $\eta_p^2 = .229$, and for the prior multiple-choice condition $F(1, 66) = 12.00$, $p < .001$, $\eta_p^2 = .154$, but not for the no prior test condition, $F(1, 69) = .40$, $p = .531$, $\eta_p^2 = .006$. This finding suggests that, regardless of JOL Type, participants in both tested conditions reduced their JOLs across sections (reducing over-confidence), while those who took no prior tests remained consistently confident in their future performance across sections. Finally, I conducted post-hoc comparisons between the Prior Review Type conditions separately for cued-recall JOLs and multiple-choice JOLs to decompose the interaction between JOL type and Prior Review Type. JOLs were higher for the no prior test condition than the prior cued-recall condition for both cued-recall JOLs (Ms = .59 and .47, respectively), $t(135) = 3.90$, $p < .001$, $d = .67$, and multiple-choice JOLs ($M = .72$ and .66, respectively), $t(135) = 2.09$, $p = .039$, $d = 0.36$ (although the effect was smaller in the multiple-choice JOLs). Participants in the prior multiple-choice condition also provided lower JOLs relative to the no prior test condition for cued-recall JOLs ($M = .48$), $t(135) = 3.23$, $p = .002$, $d = 0.56$, and multiple-choice JOLs ($M = .64$), $t(135) = 2.28$, $p = .024$, $d = 0.39$ (but again, the effect size was smaller for multiple-choice JOLs than cued-recall JOLs). Finally, there was no difference between the two tested conditions on cued-recall JOLs $t(132) = 0.29$, $p = .774$, $d = 0.05$, or multiple-choice JOLs, $t(132) = .41$, $p = .682$, $d = 0.07$.

Essentially, the results of this analysis demonstrate that prior testing (relative to no-testing) of either format reduces confidence in future performance but does so for future cued-recall performance more so that future multiple-choice performance.

Correspondence between predicted performance and actual performance for Section 4 (the criterial test, depicted in Figure 13) was calculated by subtracting actual performance from predicted performance (as a proportion). Subsequently, these scores could range from -1.0 – +1.0, with a score of -1.0 indicating total under-confidence (e.g., predicting completely inaccurate performance and later remembering all of the items), and a score of +1.0 indicating total over-confidence (e.g., predicting perfect performance, but later remembering none of the items.) All negative scores indicate under-confidence, and all positive scores indicate over-confidence. A score of zero would indicate completely accurate calibration (e.g., a participant precisely predicted their future score).

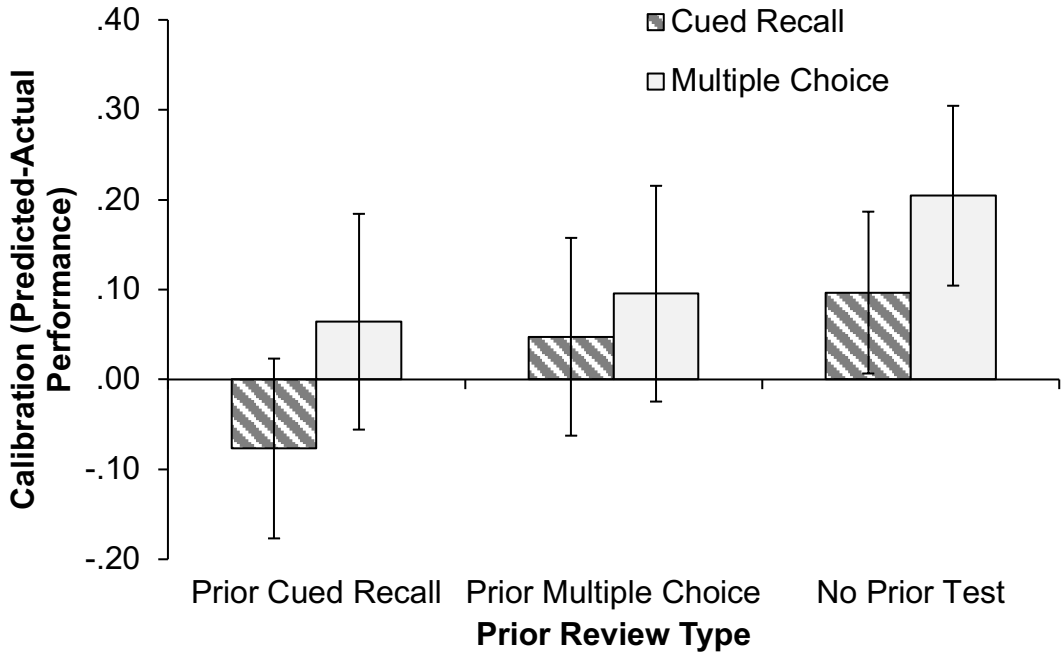

Figure 13. Over- and under-confidence as a function of prior review type and criterial test type in Experiment 2. Scores above zero indicate overconfidence, and scores below zero indicate underconfidence. Bars represent descriptive 95% confidence intervals.

For the criterial test, a 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) x 2 (Criterial Test Type: Cued-Recall or Multiple-Choice) between-subjects ANOVA yielded a main effect of Prior Review Type, $F(2, 198) = 4.53$, $p = .012$, $\eta_p^2 = .044$ and a main effect of Criterial Test Type, $F(1, 198) = 5.78$, $p = .017$, $\eta_p^2 = .028$. The interaction was not significant, $F(2, 198) = 0.30$, $p = .739$, $\eta_p^2 = .003$. In the case of the main effect of Criterial Test Type, overconfidence in future performance was higher when the criterial test was multiple-choice ($M = .13$) than when the criterial test was cued-recall ($M = .02$). For the main effect of Prior Review Type, post-hoc tests revealed that subjects in the no prior test condition ($M = .15$) were more overconfident in their performance than subjects in the prior cued-recall condition ($M = -.07$), $t(135) = 3.03$, $p = .003$, $d = 0.52$, but were not more overconfident than subjects in the prior multiple-choice condition ($M = .08$), $t(135) = 1.39$, $p = .168$, $d = 0.24$. There was no difference in scores between the two tested conditions, $t(132) = 1.54$, $p = .127$, $d = 0.27$. This suggests that experience with tests, particularly more difficult cued-recall tests, may result in reduced overconfidence.

To further examine the relation between predicted and actual scores, I conducted single-sample *t*-tests for each level of the Prior Review Type variable against the perfect calibration score of zero. In the prior cued-recall condition, the difference was not significant, $t(66) = 0.18$, $p = .855$, $d = 0.05$, indicating that calibration in this condition was near perfect (but note that this may have occurred because participants were underconfident on the criterial cued-recall test but overconfident on the criterial multiple-choice test.) However, in the prior multiple-choice condition, calibration was significantly higher than zero, $t(66) = 2.01$, $p = .048$, $d = .50$, as it was in the no prior test condition, $t(69) = 4.54$, $p < .001$, $d = 1.09$. Note that the effect size for the

no-test condition was more than double that of the multiple-choice condition, indicating greater overconfidence. Thus, individuals who took cued-recall tests prior to the criterial test were relatively well-calibrated, individuals who took multiple-choice tests were somewhat overconfident, and individuals in the no-test condition were the most overconfident in their criterial test performance.

I also examined calibration across Sections 1-3 in the two tested conditions. A 2 (Test Condition: Cued-Recall or Multiple-Choice) x 3 (Sections 1-3) mixed ANOVA showed a main effect of Section, $F(2, 264) = 11.03$, $p < .001$, $\eta_p^2 = .077$ as well as a main effect of Test Condition, $F(1, 132) = 37.81$, $p < .001$, $\eta_p^2 = .223$. In the latter case, participants in the cued-recall condition were more overconfident in their performance ($M = .23$) than participants in the multiple-choice condition ($M = .02$). In the former case, post-hoc tests demonstrated that calibration improved from Section 1 ($M = .21$) to Section 2 ($M = .08$), $t(133) = 4.30$, $p < .001$, $d = 0.37$, but remained stable from Section 2 to Section 3 ($M = .09$), $t(133) = .40$, $p = .691$, $d = 0.03$. These findings indicate that once participants in the tested conditions had reduced their confidence after the first test, they did not continue to improve on the subsequent tests.

Last, I examined the correlation between JOLs and reading times for Sections 1-4. None of these correlations were significant, r's < .07, p's > .327.

Cumulative Test Performance

Final test performance was analyzed as in Experiment 1. Overall test performance was composed of old and new items from Sections 1-4. A one-way ANOVA was performed on overall test performance (collapsing across Sections and Item Type) across the three prior test conditions, and was significant, $F(2, 202) = 11.59$, $p < .001$, $\eta_p^2 = .103$. Follow-up tests revealed backward testing effects for prior-cued recall ($M = .42$) relative to no prior test ($M = .29$), $t(136)$

= 4.83, $p < .001$, $d = 0.83$, as well as for prior multiple-choice ($M = .36$), $t(136) = 2.87$, $p = .005$, $d = 0.49$. Final test accuracy was marginally lower for prior multiple-choice relative to prior cued-recall $t(132) = 1.90$, $p = .059$, $d = 0.33$. Thus, prior cued-recall testing seemed to result in a larger backward testing effect than multiple-choice, although prior retrieval of any type was beneficial compared to no prior retrieval.

Table 5

Cumulative Test Performance as a Function of Prior Interpolated Test Condition for Old and New Items from Sections 1-3 in Experiment 2

	Section		
	1	***2***	***3***
Cued Recall			
Old	.54 *(.04)*	.51 *(.04)*	.52 *(.03)*
New	.21 *(.02)*	.24 *(.03)*	.19 *(.03)*
Multiple Choice			
Old	.49 *(.04)*	.41 *(.04)*	.40 *(.04)*
New	.20 *(.02)*	.17 *(.04)*	.13 *(.02)*
No Test			
Old	--	--	--
New	.21 *(.02)*	.17 *(.03)*	.18 *(.02)*

Note. Numbers in parentheses represent standard errors of the mean.

A second analysis was performed on the items that were previously tested or not tested in Sections 1-3 for each condition (see Table 5). In a 2 (Item Type: Previously Tested or New) x 2 (Prior Review Type: Cued Recall or Multiple Choice) mixed ANOVA, there was a main effect of Prior Review Type, such that performance was higher for participants who previously took cued-recall tests ($M = .37$) than for participants who took multiple-choice tests ($M = .30$), as in the above analysis, $F(1, 132) = 5.24$, $p = .024$, $\eta_p^2 = .038$. There was also a main effect of Item

Type, with previously-tested items (M = .48) being remembered more often than items that were not previously tested (M = .19), $F(1, 132) = 294.00$, $p < .001$, $\eta_p^2 = .690$. The interaction was not significant $F(1, 132) = .027$, $p = .243$, $\eta_p^2 = .010$. This reaffirms the finding from Experiment 1 wherein there was a within-subjects testing effect for the first three text sections.

Table 6

Cumulative Test Performance as a Function of Prior Interpolated Recall Condition for Old and New Items from Section 4 in Experiment 2

	Criterial Test Type	
Interpolated Test Type	***Cued Recall***	***Multiple Choice***
Cued Recall		
Old	.74 *(.04)*	.85 *(.04)*
New	.31 *(.04)*	.34 *(.04)*
Multiple Choice		
Old	.71 *(.05)*	.74 *(.04)*
New	.39 *(.04)*	.37 *(.05)*
No Test		
Old	.71 *(.05)*	.78 *(.04)*
New	.39 *(.05)*	.33 *(.04)*

Note. Numbers in parentheses represent standard errors of the mean.

I also examined performance for new items only in a 3 (Section 1-3) x 3 (Prior Review Type) mixed ANOVA, which showed a marginal effect of Section, $F(2, 404) = 2.94$, $p = .054$, $\eta_p^2 = .014$, but no main effect of Prior Review Type, $F(2, 202) = 1.79$, $p = .169$, $\eta_p^2 = .017$, nor an interaction, $F(4, 404) = 0.94$, $p = .442$, $\eta_p^2 = .009$. In post-hoc tests, there was no difference in performance between Section 1 (M = .20) and Section 2 (M = .19), $t(204) = 0.76$, $p = .447$, $d = 0.05$, or between Section 2 and Section 3 (M = .17), $t(204) = 1.60$, $p = .111$, $d = 0.11$, but

performance was significantly worse in Section 3 than in Section 1, $t(204) = 2.30$, $p = .023$, $d = 0.16$. This is similar to the results of Experiment 1, where new items from Section 3 appeared slightly more difficult than the other Sections.

Next, a 2 (Item Type: Previously Tested or New) x 3 (Prior Review Type: Cued Recall, Multiple Choice, or No-Test) x 2 (Criterial Test Type: Cued Recall or Multiple Choice) mixed ANOVA was performed on recall for items from the criterial test (see Table 6). This analysis yielded a main effect of Item Type. As in Sections 1-3, old items ($M = .76$) were remembered more often than new items ($M = .36$), $F(1, 199) = 347.57$, $p < .001$, $\eta_p^2 = .636$. There was no effect of Prior Review Type $F(2, 199) = 0.02$, $p = .976$, $\eta_p^2 < .001$. There was a significant interaction between Item Type and Criterial Test Type, $F(1, 199) = 3.95$, $p = .048$, $\eta_p^2 = .019$, and a marginal interaction between Item Type and Prior Review Type, $F(2, 199) = 2.20$, $p = .056$, $\eta_p^2 = .029$. The two-way interaction between Prior Review Type and Criterial Test Type was not significant, $F(2, 199) = 0.53$, $p = .592$, $\eta_p^2 = .005$, nor was the three-way interaction, $F(2, 199) = 0.31$, $p = .733$, $\eta_p^2 = .003$.

In order to decompose the significant two-way interaction between Criterial Test Type and Item Type, I examined the difference between the Criterial Test Type Conditions separately for old and new items. Old items were remembered better when the criterial test was multiple-choice ($M = .79$) than when it was cued-recall ($M = .72$), $t(203) = 2.01$, $p = .046$, $d = 0.28$, but there was no difference between the conditions when items were new ($Ms = .36$ and .35, respectively), $t(203) = .42$, $p = .673$, $d = 0.06$. Therefore, participants may have learned the feedback better when the items were tested with multiple choice than when tested with cued recall. Finally, I examined the marginal two-way interaction between Item Type and Prior Review Type. I compared the difference between old and new items (i.e., the within-subjects

testing effect) separately for each level of prior review. In the cued recall condition, there was a very large difference between old items (M = .80) and new items (M = .33), $t(66) = 13.40$, $p < .001$, $d = 1.63$. The difference was somewhat smaller (but still large) for the multiple-choice condition (Ms = .73 and .38 for old and new items), $t(66) = 9.32$, $p < .001$, $d = 1.14$. In the no-prior test condition, the difference was similar to that as the multiple-choice condition (Ms = .74 and .36), $t(70) = 9.72$, $p < .001$, $d = 1.15$. Thus, the marginal interaction seems to be driven by the fact that there was a larger within-subjects testing effect on the criterial test when participants were previously tested with cued recall relative to multiple choice or no prior test.

Test Expectancy

As in Experiment 1, participants were asked to indicate which activity they had expected after the reading of the final section of the passage (see Table 7). In general, it appears that both prior test conditions expected a cued-recall test for Section 4 more often than participants who had no prior test, and participants who had not been previously tested disproportionately reported a multiple-choice expectation.

Table 7

Section 4 Test Expectations as a Percentage of Each Prior Test Condition

	Expectation		
Prior Review Type	***Cued Recall***	***Multiple Choice***	***No Test***
Cued Recall	58%	32%	9%
Multiple Choice	49%	37%	14%
No Test	26%	50%	24%

Discussion

Unlike Experiment 1, Experiment 2 did not provide substantial support for the metacognitive account for the forward effect of testing. Importantly, no forward effect of testing was observed regardless of prior test condition or criterial test condition. Referring to the former finding, when comparing the means for the criterial test between Experiments 1 and 2, it appears that this lack of an effect is due primarily to an increase in performance in the no-test condition (M = .38 in Experiment 1 vs. .52 in Experiment 2). It is not uncommon for judgments of learning to eliminate the benefits of testing by increasing performance in non-tested conditions (e.g., Dougherty, Scheck, Nelson, and Narens, 2005), although this is the first demonstration of this effect as it applies to the forward effect of testing. It is possible that querying judgments of learning four times per section may have increased test expectancy or may have encourage participants to reflect on their learning and change strategies. This may have led participants to pay more attention during passage reading, an idea that is bolstered by the reading time analysis. Whereas reading times fell in the no-test condition (compared to the tested conditions) in Experiment 1, they remained stable and consistent with reading times in the tested conditions in Experiment 2. However, the finding that there was no main effect of criterial test type (e.g., performance was not better on the criterial multiple-choice test than the criterial cued-recall test) is more difficult to explain. It may be the case that providing JOLs slightly increased test expectancy in the tested conditions as well as the non-tested condition, improving performance. Regardless, requiring participants to reflect on their future performance clearly had a positive impact on their learning, particularly in the absence of interpolated testing.

However, despite the null effects of the manipulations on accuracy, there were differences between conditions on explicit metacognitive judgments. First, individuals in the no-

prior test condition tended to report higher JOLs across sections than those who received interpolated tests. Those who were tested demonstrated drops in JOLs after the first section (a reduction of overconfidence effect). Furthermore, tested individuals were better calibrated (i.e., their predictions of future performance more closely matched actual performance), and this was more so following interpolated cued-recall than multiple-choice testing. However, it is important to note again that these differences in metacognitive judgments did not translate into actual performance differences.

Whereas testing did not produce a forward effect in Experiment 2, it did produce powerful backward testing effects on the cumulative test both between-subjects (i.e., comparing the tested conditions to the non-tested condition) and within-subjects (i.e., comparing old to new items). Therefore, it remains clear that retrieval practice has a large advantage over no retrieval practice on learning, even when short-term consequences of providing metacognitive judgments might have occluded the beneficial forward effect.

One question that remains is how one might make multiple-choice testing more effective at enhancing learning (or how one might design multiple-choice tests to result in better calibration). One way to do this is to reduce the overconfidence that is associated with recognition fluency. In order to do this, one would need to make multiple-choice tests more difficult. In Experiment 3, I did so by increasing the plausibility or competitiveness of the lures on the multiple-choice tests, which should reduce over-confidence and increase calibration.

CHAPTER 4. EXPERIMENT 3

Method

Participants, Design, Materials, and Procedure

Three hundred and twenty-nine participants were recruited from Iowa State University and received partial course credit for their participation. Twenty-four participants were eliminated from the analysis for not finishing the experiment, four were eliminated because they indicated that they had read the materials before, four were eliminated for leaving the experiment for a duration of greater than 10 minutes, and 15 were eliminated for indicating that English was not their native language. This yielded 282 participants for the final analysis. The number of participants per between-subjects condition is displayed in Table 1.

Experiment 3 used a 2 (Criterial Test Type: Cued Recall or Multiple Choice) x 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) between-subjects design, and the primary dependent variable of interest was Section 4 (criterial) test performance.

The procedure was similar to Experiment 2 but with an additional condition. Participants in the Easy Multiple-Choice condition completed the same multiple-choice tests for Sections 1-3 as in the Multiple-Choice condition in Experiments 1 and 2. However, Experiment 3 also included a Difficult Multiple-Choice condition. Participants in this group received the same question stems as the Easy Multiple-Choice condition but received lures that were more competitive or plausible than those used in Experiments 1 and 2. This manipulation makes answering these questions more difficult. For example, in the question: "Who invented the laser?" the correct answer is Gordon Gould. In the Easy Multiple-Choice condition, participants received the three lures: Theodore Maiman, Albert Einstein, and Stephen Hawking. In the

Difficult Multiple-Choice condition, participants received the three lures: Geoffrey Gould, Abraham Gould, and Gordon Maiman. These lures were designed to be competitive in the sense that it would make picking the correct answer more difficult without explicit retrieval (i.e., recall) of the correct answer. This procedure was used because Little, Bjork, and Bjork, (2012) found that highly competitive (i.e., plausible) lures can be very effective in enhancing the backward testing effect when recognition is used as the initial test, even to the same degree or greater than cued-recall questions. In a similar manner, difficult multiple-choice questions may make participants less confident and therefore more likely to adopt better encoding strategies across sections. Importantly, the multiple-choice test for Section 4 was the same as in Experiments 1 and 2, making this test section more similar to the Easy Multiple-Choice questions from Sections 1-3.

Results

Interpolated and Criterial Test Performance

A 3 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, or Difficult Multiple-Choice) x 3 (Section 1-3) mixed ANOVA was used to examine interpolated test performance (see Figure 14). There was a main effect of Section, $F(2, 416) = 4.92$, $p = .008$, $\eta_p^2 = .023$, as well as a main effect of Prior Review Type, $F(2, 208) = 68.10$, $p < .001$, $\eta_p^2 = .396$. The interaction was not significant, $F(4, 416) = 1.17$, $p = .322$, $\eta_p^2 = .011$. Overall, performance was not different between Section 1 ($M = .51$) and Section 2 ($M = .52$), $t(210) = .43$, $p = .671$, $d = 0.03$, but performance was lower in Section 3 ($M = .46$) than in Section 2, $t(210) = 2.85$, $p = .005$, $d = 0.20$, and Section 1 $t(210) = 2.50$, $p = .013$, $d = 0.17$. However, note that this difference in performance from Section 3 compared to Sections 1 and 2 was small (ds < 0.21). The cued-recall tests ($M = .21$) were also more difficult than the easy multiple-choice tests ($M = .48$),

$t(135) = 11.42$, $p < .001$, $d = 1.97$ and difficult multiple-choice tests ($M = .41$), $t(139) = 8.48$, $p < .001$, $d = 1.44$. Performance was also worse in the difficult multiple-choice condition than in the easy multiple-choice condition, $t(142) = 2.88$, $p = .005$, $d = 0.48$. Therefore, my manipulation of lure similarity was successful in reducing performance in the prior difficult multiple-choice condition.

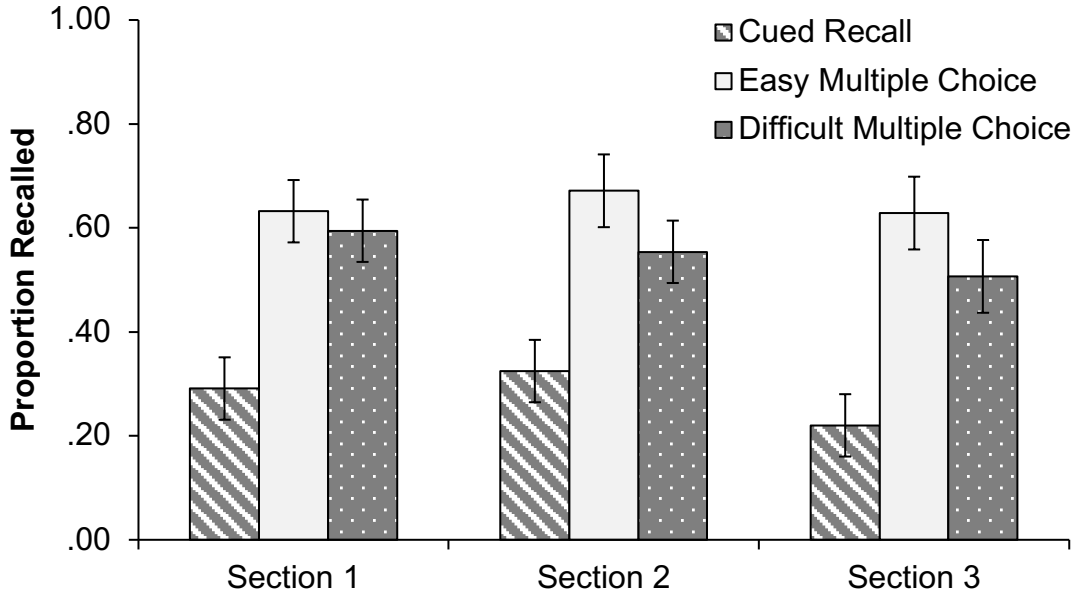

Figure 14. Interpolated test performance for Sections 1-3 as a function of interpolated test type in Experiment 3. Bars represent descriptive 95% confidence intervals.

A 2 (Criterial Test Type: Cued Recall or Multiple Choice) x 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) between-subjects ANOVA was used to examine performance on the criterial test (see Figure 15). As in Experiment 2, there was no main effect of Prior Review Type $F(3, 274) = 0.11$, $p = .953$, $\eta_p^2 = .001$, no main effect of Criterial Test Type, $F(1, 274) = 0.26$, $p = .261$, $\eta_p^2 = .005$, nor was there an interaction, $F(3, 274) = 1.61$, $p = .187$, $\eta_p^2 = .017$. This signifies again that when JOLs were required, there was no observed forward effect of testing, nor was there an effect of test format on the criterial test. As

in Experiment 2, this finding is surprising. I will return to these null effects in the General Discussion.

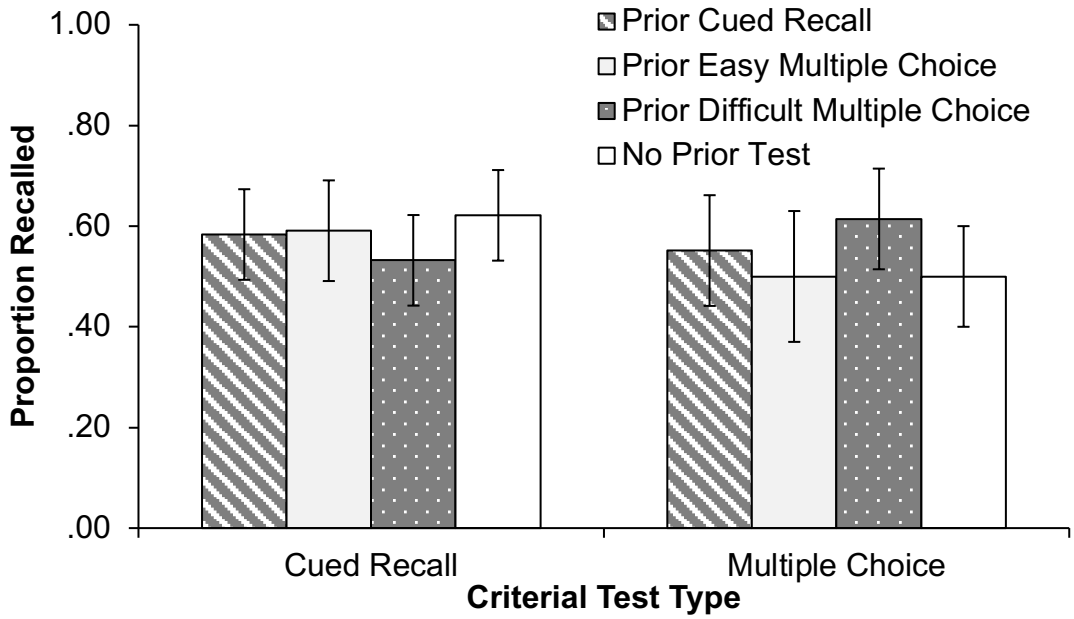

Figure 15. Criterial test performance for Section 4 as a function of prior interpolated test type in Experiment 3. Bars represent descriptive 95% confidence intervals.

Section 1-4 Reading Times

A 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) x 4 (Section: 1-4) mixed ANOVA on reading times (see Figure 16) demonstrated that there was no main effect of Section, $F(3, 834) = 0.66$, $p = .579$, $\eta_p^2 = .002$, no main effect of Prior Review Type, $F(3, 278) = 1.27$, $p = .284$, $\eta_p^2 = .014$, nor was there an interaction, $F(9, 834) = 1.22$, $p = .277$, $\eta_p^2 = .013$. Thus, as in Experiment 2, there was no effect of interpolated testing on reading times. Linear contrasts confirmed this assertion, with no linear trend observed for the prior cued-recall condition, $F(1, 66) = 0.70$, $p = .407$, $\eta_p^2 = .010$, the prior easy multiple-choice condition, $F(1, 69) = 1.44$, $p = .234$, $\eta_p^2 = .020$, the prior difficult multiple-choice condition, $F(1,$

73) = 0.31, p = .580, η_p^2 = .004, nor the no prior test condition, F(1, 70) = 2.67, p = .107, η_p^2 = .037.

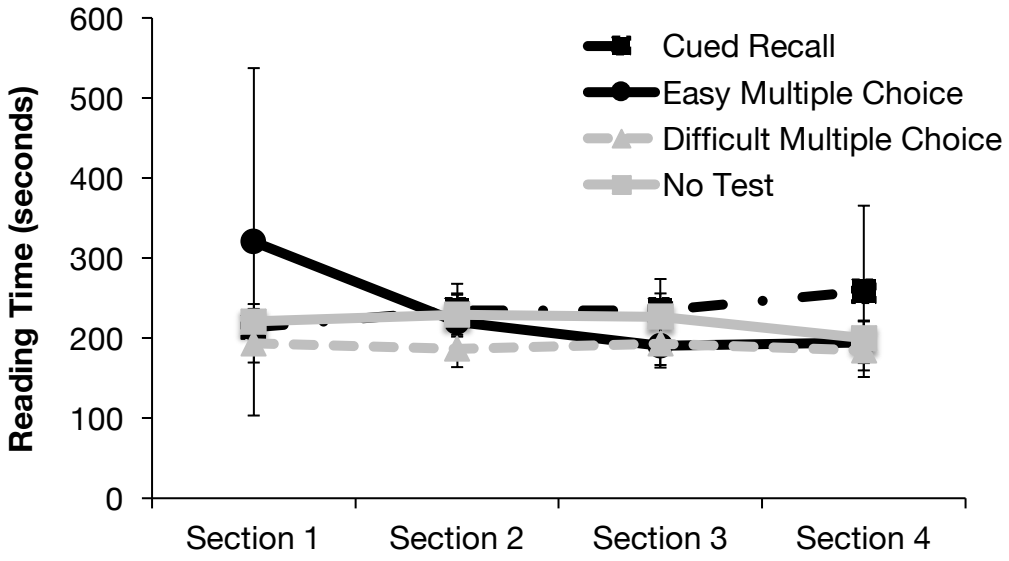

Figure 16. Reading time as function of prior interpolated test condition for Sections 1-4 in Experiment 3. Bars represent descriptive 95% confidence intervals.

I also examined the correlation between Section 4 reading time and criterial test performance. When the criterial test was cued recall, there was a significant correlation, r = .19, p = .024, but there was not a significant relationship when the criterial test was multiple-choice, r = .03, p = .709.

Judgments of Learning

Judgments of learning across sections as a function of prior review type are depicted in Figure 17. As in Experiment 2, I analyzed JOLs first by performing a 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) x 4 (Section: 1-4) x 2 (JOL Test Type: Cued-Recall or Multiple-Choice) mixed ANOVA. There was a main effect of Section, F(3, 834) = 48.66, p < .001, η_p^2 = .149, and a main effect of JOL Test Type, F(1, 278) = 505.91, p < .001, η_p^2 = .645. There was no main effect of Prior Review Type, F(3, 278) = 1.51,

p = .213, η_p^2 = .016. None of the two-way interactions were significant, including that between Section and Prior Review Type, $F(9, 834) = 1.24$, p = .265, η_p^2 = .013, JOL Test Type and Prior Review Type, $F(3, 278) = 1.39$, p = .247, η_p^2 = .015, and Section and JOL Test Type, $F(3, 834)$ = 1.78, p = .150, η_p^2 = .006. The three-way interaction was also not significant, $F(9, 834) = 1.28$, p = .247, η_p^2 = .014. The main effect of JOL Test Type indicates that participants made higher JOL estimates for multiple-choice tests (M = .67) than cued-recall tests (M = .52), suggesting that individuals are aware of the greater difficulty of cued-recall tests.

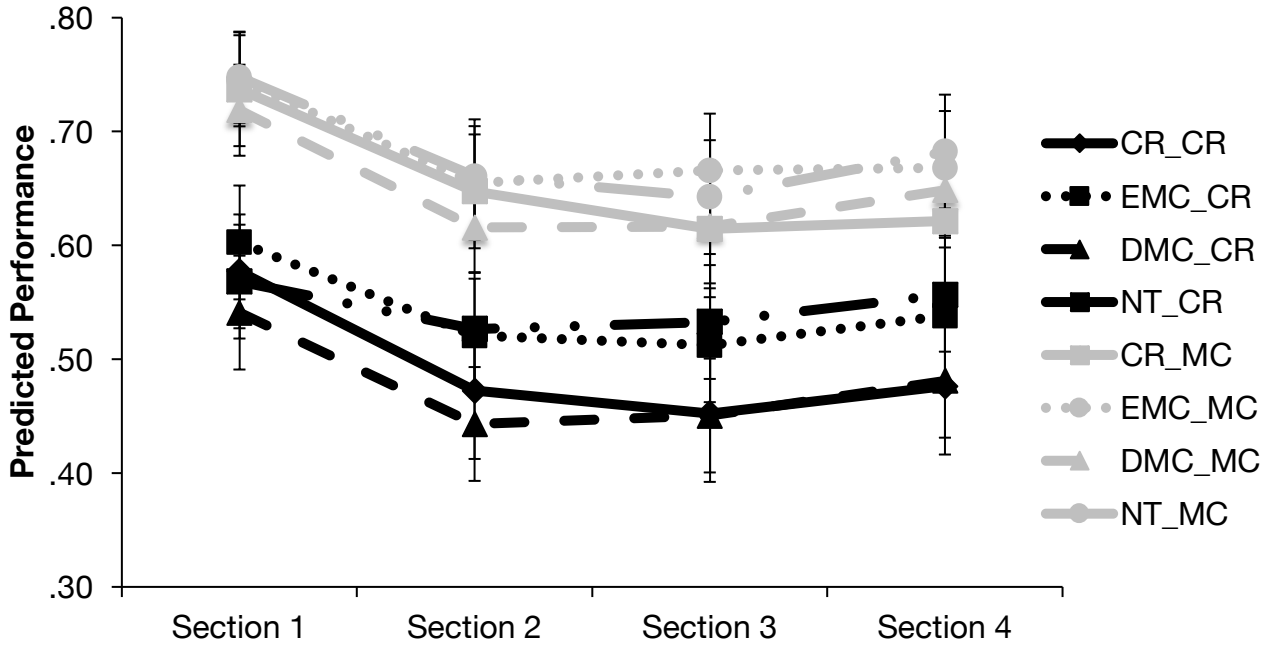

Figure 17. JOL's provided by subjects for Sections 1-4 in Experiment 3 as a function of prior testing and type of test prediction. The first denotation in each condition refers to the prior test condition (CR = cued-recall, EMC = easy multiple-choice, DMC = difficult multiple-choice, and NT = no test), and the second denotation refers to the type of test prediction. Gray lines indicate predictions made regarding future multiple-choice performance, while black lines indicate predictions made regarding future cued-recall performance.

As can be seen in Figure 17, the main effect of Section is due to JOL's decreasing across sections, as confirmed by the linear contrast $F(1, 278) = 43.67$, $p < .001$, η_p^2 = .136. Post-hoc tests showed that JOLs declined from Section 1 (M = .65) to Section 2 (M = .57), $t(281) = 9.82$, p

< .001, d = 0.59, remained stable from Section 2 to Section 3 (M = .56), $t(281) = 0.82$, p = .412, d = 0.05, and rose slightly from Section 3 to Section 4 (M = .58), $t(281) = 3.21$, p = .001, d = 0.19. This finding demonstrates that learners adjust their judgments of learning after a single learning experience, after which they remain approximately stable, although this must be interpreted cautiously as each section contains different material.

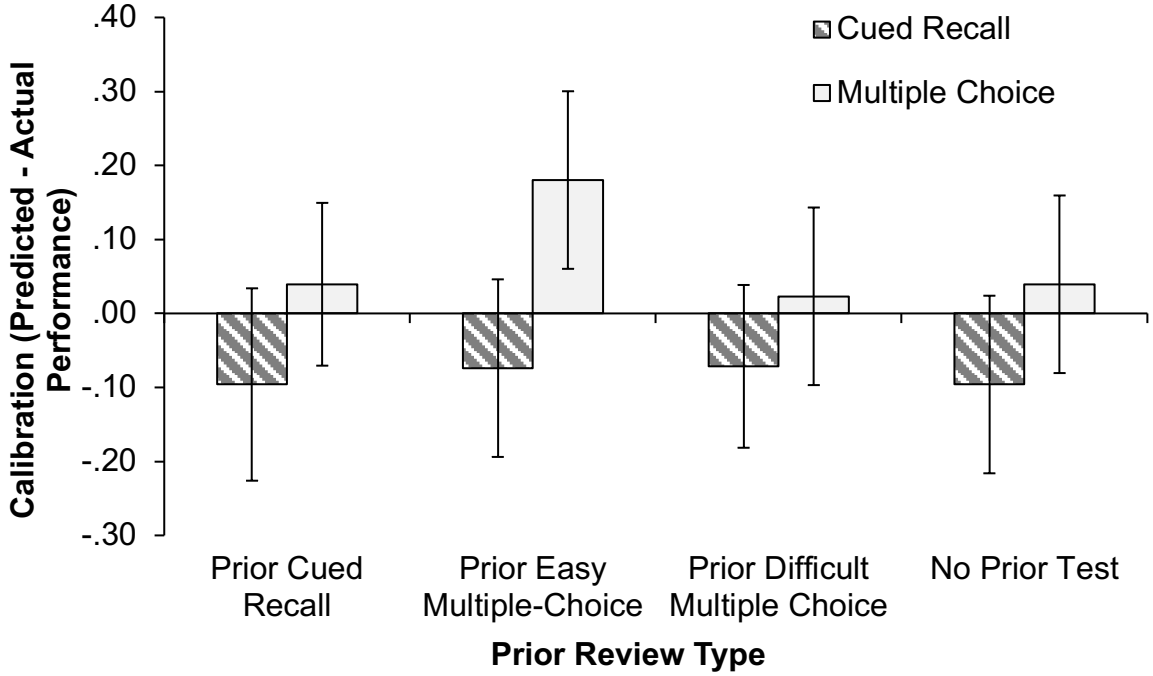

Figure 18. Over- and under-confidence as a function of prior review type and criterial test type in Experiment 3.

Calibration (see Figure 18) for Section 4 was analyzed in a 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) x 2 (Criterial Test Type: Cued-Recall or Multiple-Choice) between-subjects ANOVA. This test yielded a main effect of Criterial Test Type, $F(1, 274) = 23.46$, $p < .001$, $\eta_p^2 = .079$, showing that learners tended to be underconfident on the criterial cued-recall test, but overconfident on the criterial multiple-choice test. I separately tested the calibration scores for each condition against the perfect calibration score of zero, and found that learners were significantly underconfident on the criterial cued-

recall (M = -.08) test, $t(139) = 2.87$, $p = .005$, $d = 0.49$. On the criterial multiple-choice test (M = .12), learners were significantly overconfident, $t(141) = 4.02$, $p < .001$, $d = 0.68$. The main effect of Prior Review Type was not significant, $F(3, 274) = 1.40$, $p = .244$, $\eta_p^2 = .015$, nor was the interaction between the two variables, $F(3, 274) = 1.46$, $p = .225$, $\eta_p^2 = .016$.

I also examined calibration scores for the three tested conditions across Sections 1-3. A 3 (Prior Test Condition: Cued-Recall, Easy Multiple-Choice, or Difficult Multiple-Choice) x 3 (Sections 1-3) mixed ANOVA yielded a main effect of Section, $F(2, 416) = 6.91$, $p = .001$, $\eta_p^2 = .032$, and a main effect of Test Condition, $F(2, 208) = 5.28$, $p = .006$, $\eta_p^2 = .048$, but not an interaction, $F(4, 416) = 1.75$, $p = .138$, $\eta_p^2 = .017$. Post-hoc tests revealed overconfidence was reduced slightly from Section 1 (M = .21) to Section 2 (M = .12), $t(210) = 3.61$, $p < .001$, $d = 0.25$, but increased slightly from Section 2 to Section 3 (M = .16), $t(210) = 2.10$, $p = .037$, $d = .14$. Essentially, participants became less overconfident overall from the first to second section, but became slightly more overconfident from the second to third section. Additionally, participants in the prior cued-recall condition (M = .07) were less overconfident than participants in the prior easy multiple-choice condition (M = .23), $t(135) = 2.94$, $p = .004$, $d = 0.51$, and participants in the prior difficult multiple-choice condition (M = .19), $t(139) = 2.65$, $p = .009$, $d = 0.45$. There was no difference between the two prior multiple-choice conditions, $t(142) = 0.86$, $p = .394$, $d = 0.14$.

Last, I examined the correlation between JOLs from Sections 1-4 and reading times for Section 4. None of these correlations were significant, r's < .10, ps > .12.

Cumulative Test Performance

As in Experiments 1 and 2, a one-way ANOVA comparing overall test performance across the Prior Review Type conditions was conducted. Overall test performance was composed

of old and new items from Sections 1-4. The effect of Prior Test Condition was significant, $F(3, 279) = 13.19$, $p < .001$, $\eta_p^2 = .124$. There was a backward testing effect for prior cued recall ($M = .45$) relative to no prior test ($M = .29$), $t(133) = 6.06$, $p < .001$, $d = 1.05$, for prior easy multiple-choice ($M = .43$), $t(139) = 5.48$, $p < .001$, $d = 0.93$, and for prior difficult multiple-choice ($M = .40$), $t(147) = 4.58$, $p < .001$, $d = 0.75$. There was no difference in final test performance between the prior cued-recall condition and prior easy multiple-choice test condition, $t(132) = 0.62$, $p = .538$, $d = 0.11$, or prior difficult multiple-choice condition, $t(140) = 1.58$, $p = .118$, $d = 0.27$. The two prior multiple-choice conditions also did not differ from each other, $t(146) = 0.96$, $p = .339$, $d = 0.16$. Thus, testing of any type resulted in a backward retrieval practice effect.

Table 8

Cumulative Test Performance as a Function of Prior Interpolated Test Condition for Old and New Items from Sections 1-3 in Experiment 3

	Section		
	1	***2***	***3***
Cued Recall			
Old	.52 *(.04)*	.58 *(.04)*	.52 *(.04)*
New	.19 *(.02)*	.26 *(.03)*	.27 *(.03)*
Easy Multiple Choice			
Old	.59 *(.03)*	.50 *(.04)*	.53 *(.03)*
New	.20 *(.02)*	.24 *(.03)*	.23 *(.03)*
Difficult Multiple Choice			
Old	.50 *(.03)*	.45 *(.03)*	.48 *(.03)*
New	.23 *(.03)*	.24 *(.03)*	.19 *(.03)*
No Test			
Old	--	--	--
New	.23 *(.02)*	.18 *(.03)*	.16 *(.02)*

Note. Numbers in parentheses represent standard errors of the mean.

Next, I examined the difference between new and old items from Sections 1-3 (see Table 8) in a 2 (Item Type: Previously Tested or New) x 3 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, or Difficult Multiple-Choice) mixed ANOVA. There was a main effect of Item Type, with old items (M = .52) being remembered more often than new items (M = .23), F(1, 212) = 522.27, p < .001, η_p^2 = .711. The main effect of Prior Review Type was not significant, F(2, 212) = .939, p = .393, η_p^2 = .009, nor was the interaction, F(2, 212) = 1.59, p = .207, η_p^2 = .015. Thus, there was a large within-subjects testing effect that did not depend on the type of review for Sections 1-3.

I also examined new items separately in a 3 (Section: 1-3) x 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, or Difficult Multiple-Choice) mixed ANOVA. There was no main effect of Section, F(2, 564) = 1.39, p = .249, η_p^2 = .005, nor was there a main effect of Prior Review Type, F(3, 282) = 0.82, p = .486, η_p^2 = .009. However, there was an interaction, F(6, 564) = 2.75, p = .012, η_p^2 = .028. To decompose the interaction, I ran four separate one-way within-subjects ANOVAs for each prior review condition. The main effect of Section was significant for those who had taken prior cued-recall tests, F(2, 132) = 4.73, p = .010, η_p^2 = .067, and was marginal for those who had taken no prior tests, F(2, 140) = 2.68, p = .072, η_p^2 = .037. It was not significant for either the prior easy multiple-choice condition, F(2, 138) = 0.78, p = .463, ηp2 = .011, or the prior difficult multiple-choice condition, F(2, 154) = 1.90, p = .154, η_p^2 = .024. In the prior cued-recall condition, post-hoc tests determined that performance was lower for Section 1 (M = .19) than Section 2 (M = .26), t(66) = 2.70, p = .009, d = 0.33, and Section 3 (M = .27), t(66) = 2.60, p = .012, d = 0.32. There was no difference in performance between Sections 2 and 3, t(66) = 0.13, p = .896, d = 0.01. In the no prior test condition, there was no difference in performance between Sections 1 (M = .23) and 2 (M = .18), t(70) = 1.40, p = .165, d

= 0.17, and no difference in performance between Sections 2 and 3 (M = .17), $t(70) = 0.79$, p = .435, d = 0.09, but performance was higher in Section 1 than in Section 3, $t(70) = 2.48$, p = .016, d = 0.29. Therefore, which section had lower performance depended on prior review condition.

Table 9

Cumulative Test Performance as a Function of Prior Interpolated Recall Condition for Old and New Items from Section 4 in Experiment 3

	Criterial Test Type	
Interpolated Test Type	***Cued Recall***	***Multiple Choice***
Cued Recall		
Old	.84 *(.03)*	.80 *(.03)*
New	.42 *(.05)*	.37 *(.04)*
Easy Multiple Choice		
Old	.80 *(.03)*	.72 *(.03)*
New	.38 *(.05)*	.42 *(.04)*
Difficult Multiple Choice		
Old	.77 *(.03)*	.82 *(.03)*
New	.38 *(.04)*	.33 *(.04)*
No Test		
Old	.81 *(.03)*	.76 *(.03)*
New	.36 *(.04)*	.33 *(.04)*

Note. Numbers in parentheses represent standard errors of the mean.

I analyzed items that were from the criterial section in a 2 (Item Type: Previously Tested or New) x 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) x 2 (Criterial Test Type: Cued Recall or Multiple Choice) mixed ANOVA, and these data are presented in Table 9. There was a main effect of Item Type, such that previously tested items (M = .79) were remembered more often than new items (M = .39), $F(1, 278) = 487.19$, $p <$

.001, $\eta_p^2 = .637$. There was no effect of Prior Review Type, $F(1, 278) = 0.83$, $p = .480$, $\eta_p^2 = .009$, nor was there a main effect of Criterial Test Type, $F(1, 278) = 0.52$, $p = .470$, $\eta_p^2 = .002$. The two way interactions were not significant, including that between Item Type and Prior Review Type, $F(3, 278) = 0.97$, $p = .407$, $\eta_p^2 = .010$, Item Type and Criterial Test Type, $F(1, 278) = 0.51$, $p = .476$, $\eta_p^2 = .002$, and Prior Review Type and Criterial Test Type, $F(3, 278) = 0.59$, $p = .619$, $\eta_p^2 = .006$. The three-way interaction was marginal, $F(3, 278) = 2.16$, $p = .093$, $\eta_p^2 = .023$. Ultimately, this analysis affirms that prior testing of items resulted in better memory than when items were not previously tested.

Test Expectancy

Again, participants reported which activity they expected after the criterial section was read (but not tested). There was a trend towards those in the cued recall condition expecting a cued-recall test over the other options, and the multiple-choice conditions expected a multiple-choice test as well. The no prior test condition again demonstrated a bias toward the multiple-choice test as well.

Table 10

Section 4 Test Expectations as a Percentage of Each Prior Test Condition

	Expectation		
Prior Review Condition	***Cued Recall***	***Multiple Choice***	***No Test***
Cued Recall	49%	37%	13%
Easy Multiple-Choice	32%	55%	12%
Difficult Multiple-Choice	36%	48%	16%
No Test	30%	48%	23%

Discussion

Experiment 3 largely replicated the results of Experiment 2. As in Experiment 2, Experiment 3 demonstrated that, when learners are required to reflect on their future performance, there is no forward effect of testing. This may be due to an increase in test expectancy, particularly in the no-test group. This idea was supported by the finding that reading time remained stable across sections, rather than declining in the no-test condition. Furthermore, there was no main effect of criterial test. Again, this is surprising as one would expect higher performance on the criterial multiple-choice test than the criterial cued-recall test.

Curiously, the findings from the explicit judgments of learning did not entirely replicate from Experiment 2. There was a drop in JOLs across sections, indicating that learners become less confident in their future performance. Learners also predicted better performance for multiple-choice tests than cued-recall tests. However, there was no difference in JOLs based on prior test conditions, although there was a numerical trend towards cued-recall and difficult multiple-choice tests showing lower JOLs than the no-test and easy multiple-choice conditions. I will refer again to this finding in the General Discussion.

Again, despite the lack of a *forward* testing effect, there was still evidence that retrieval practice was a powerful memory enhancer. There were large backward testing effects for both between-subjects (comparing the tested conditions to the no-test condition) and within-subjects (comparing previously-tested to new items) comparisons. Therefore, while requiring evaluative judgments in lieu of a test may prevent the forward effect of testing from occurring, they do not enhance original learning on a later cumulative test.

CHAPTER 5. GENERAL DISCUSSION

In three experiments, I tested the metacognitive account of the forward effect of testing. In Experiment 1, I manipulated prior test difficulty by providing participants with either cued-recall or multiple-choice questions following the reading of each text section. Critically, successful performance on a cued-recall test relies much more heavily on the more resource-intensive process of recollection, while multiple-choice can rely on recollection, familiarity, or both. I compared both of these conditions to a no prior test condition, in which participants simply proceeded from each passage to the next with no interpolated task. I also manipulated the format of the criterial test, such that the test for the final section was either cued-recall or multiple-choice. First, I expected that prior testing would enhance new learning, regardless of the prior test format. However, I expected that this benefit of testing would depend on the format of the three previous section tests and the format of the criterial test. Namely, I predicted that prior multiple-choice testing would lead to a reduced benefit (relative to prior cued-recall testing) when the final section recall test was cued recall. This prediction arose from the finding in the backward testing effect literature that recognition tests usually result in a smaller (or negligible) testing effect than more effortful retrieval, such as recall (e.g., Chan & McDermott, 2007). Because successful recognition can rely on either recollection or familiarity (or both), performance on recognition tests is viewed as relatively easier than free recall or cued recall. However, when the criterial test was multiple-choice, I predicted that there would be no difference between the two previously tested conditions. Recognition tests are often less sensitive to differences in encoding, which could eliminate any beneficial effect of more effortful retrieval.

In Experiment 1, prior cued-recall testing resulted in a larger forward testing effect (when compared to the no-prior test condition) than prior multiple-choice testing, regardless of criterial test format. Additionally, performance was better on the criterial multiple-choice test than the criterial cued-recall test (an effect that reflects the relative ease of multiple-choice tests compared to cued-recall). However, prior cued-recall testing produced a forward testing effect that was nearly double the size (d = 0.59) of prior multiple-choice testing (d = 0.33). The fact that this effect persisted even when the criterial test was multiple-choice test is somewhat surprising and represents a novel contribution to the literature. Given that individuals who took prior multiple-choice tests should have been well practiced at the multiple-choice test format, one would have expected equivalent recall between the two testing groups. In fact, even though they had never answered a multiple-choice question for the studied materials, individuals who took prior cued-recall tests outperformed those who took prior multiple-choice tests. While surprising, this finding dovetails nicely with work examining the backward testing effect. In this literature, more effortful retrieval (e.g., recall relative to recognition) frequently produces a larger benefit of retrieval practice, regardless of final test format (Carpenter & Delosh, 2006).

Despite the fact that prior multiple-choice testing was not as potent as cued-recall at producing a forward testing effect, both types of retrieval had an impact on reading times relative to no prior testing. In the no-prior test condition, reading times dropped considerably from Section 1 to Section 2, and demonstrated a negative linear pattern across all four sections. In both testing conditions, reading time remained stable across sections. While testing did not increase reading times per se, it does seem that prior testing (regardless of the test format) discouraged participants from *reducing* their reading times across sections, similar to Yang et al (2017).

This is also consistent with work by Szpunar et al. (2013), who found that testing reduced inattention (indexed by mind wandering) during a video lecture. This finding also supports the account that suggests that testing encourages what I termed in the Introduction to be "good student" behaviors. These include reduced mind wandering, increased note taking, and increased attendance (Jing et al., 2016; Szpunar et al., 2013, Szpunar et al, 2014). More time spent reading also falls under this umbrella of positive behaviors that occur as a result of prior testing. Therefore, whereas prior multiple-choice testing did not lead to as large of a memorial benefit on learning as prior-cued recall testing, it did have an effect on behavior. I interpret the lack of reduction in reading times in the previously tested conditions to reflect perhaps an implicit metacognitive judgment regarding how much effort is required to learn the text. In both tested conditions, participants may have realized that sustained attention was needed in order to perform successfully on upcoming tests. However, the memorial benefit for cued-recall over multiple-choice may reflect qualitative changes in encoding strategy (e.g., using imagery, connecting information to one's own life, etc.) that are not reflected in reading times.

Although I interpreted reading times as an *indirect* measure of metacognitive judgments, I required explicit judgments of learning (JOLs) in Experiments 2 and 3. One major aim of this dissertation was to examine how these explicit judgments would be affected by prior testing. I predicted reduced JOL's, and better calibration between predicted and actual performance for both testing conditions relative to no prior testing. I further predicted that calibration should be better following cued-recall than multiple-choice tests. This is an important finding for the classroom, as predicted judgments of learning can influence how students engage in further study (Dunlosky & Rawson, 2012). Thus, understanding how testing influences predictions of future performance can provide important insight into the mechanisms of the forward testing effect.

Recall that in Experiments 2 and 3, participants provided four judgments of learning: one for cued-recall performance immediately, one for multiple-choice performance immediately, one for cued-recall performance on a delayed test, and one for multiple-choice performance on a delayed test. I will first discuss accuracy on the criterial test in these Experiments. Surprisingly, prior testing (cued-recall and multiple-choice in Experiment 2, and cued-recall, easy multiple-choice, and difficult multiple-choice in Experiment 3) did not result in a forward testing effect relative to no prior testing on the criterial test. The primary difference between Experiment 1 and Experiments 2 and 3 was the presence of JOLs. Therefore it seems likely that the reason for the null effect is that JOLs provided some sort of metacognitive benefit that enhanced learning, particularly among the participants who had not been previously tested. In fact, performance in Experiment 1 for the no-test condition (regardless of criterial test format) was .39, but was .52 in Experiment 2, and .56 in Experiment 3. Note that this represents a cross-experimental comparison, and caution should be exercised when making comparisons across experiments (as participants were not randomly assigned to experiments). However, this provides some preliminary evidence that the lack of the forward testing effect was not the result of reduced performance in the experimental conditions (e.g., prior cued-recall and prior multiple-choice testing), but was the result of *increased* performance in the control condition.

The finding that providing JOLs can influence performance in a control condition is not new. There is a wealth of research (for a review see Rhodes, 2016) showing that JOLs can minimize the effect of retrieval practice as well as influence future behavior. One primary way that JOLs can obfuscate the beneficial backward effect of retrieval practice is that item-by-item JOLs may induce covert retrieval. For example, a learner may study the paired-associate "green-frog," and then later be asked to predict their future performance on the item given the word

"green." Some participants may covertly retrieve the answer "frog" in order to give their JOL. This essentially turns a no-testing control into a retrieval condition, improving later final test performance. The likelihood that this may have occurred is not high Experiments 2 and 3, as I required *aggregate* JOLs (e.g., asking to reflect on performance for the entire section, rather than on an item by item basis).

Though the increase in performance for non-tested participants might not be explained by covert retrieval in Experiments 2 and 3, there are two potential alternatives that might account for this finding. First, providing JOLs might have led participants to evaluate and change their future encoding strategies. Sahakyan et al. (2004) found that providing evaluative judgments (like JOLs) prior to learning a new set of materials enhanced learning of that list (even in the presence of an instruction to forget and relative to when an evaluative judgment was not required). The authors theorized that providing this judgment might have led participants to evaluate and change their subsequent encoding strategy. Rhodes (2016) has termed this phenomenon "reactivity," in which participants may react differently to future learning opportunities after providing a JOL than when a JOL is not required. However, the research has been largely mixed, with one recent paper suggesting that there is "little evidence for a claim that metacognitive judgments enhance later retrieval," (Dougherty, Robey, & Buttaccio, 2018, p. 564). However, it is important to note that the paradigm on which Dougherty et al. based this claim is much different from the one used in the present studies and required all participants to engage in retrieval before making the item-by-item JOLs.

Another potential mechanism by which JOLs could influence future memory performance is by increasing test expectancy. The idea here is that, particularly in the no-prior test condition, learners who are repeatedly *not* tested begin to expect a test less, which then

affects how well they encode subsequent sections. One study by Weinstein, Gilmore, McDermott, and Szpunar (2014) provides support for this idea. These studies used a paradigm similar to Szpunar et al. (2008), which involves the learning of multiple word lists. Notably, after each list, subjects were either tested or not, and were asked to indicate how likely it was that they would receive a test for that list. For tested participants, test expectancy increased across lists. In contrast, for non-tested participants, test expectancy decreased across lists. Furthermore, when non-tested participants were warned that they would receive a test, they increased their judgment of how likely a test was to occur, which also increased their criterial test performance. This resulted in a reduction in the magnitude of the forward testing effect. In the case of Experiments 2 and 3, the JOLs may have served as an indirect warning or reminder that a future test was upcoming, which then increased performance overall. However, this hypothesis is not supported by the retrospective reports of expectancy. In Experiment 1 (when JOLs were not provided), 76% of participants in the no prior test condition reported expecting some sort of test (i.e., either cued-recall or multiple-choice). In Experiments 2 and 3, 76% and 77% of participants in the no-prior test conditions also reported expecting a test. However, it is important to note that these retrospective judgments may not reflect what participants may have expected prior to the criterial test. The judgments occurred long after the criterial test (after the distractor tasks and the cumulative test). Therefore, despite these data, I argue that the expectancy hypothesis may still be a viable explanation for how JOLs increased performance in the no prior test conditions in Experiments 2 and 3.

It is also important to note that the notion of JOLs inducing reactivity (Rhodes, 2016) or expectancy (Weinstein et al., 2014) may represent different metacognitive effects. Rhodes has argued that reactivity occurs when providing a JOL influences learners to adopt more effective

learning strategies after the first learning opportunity, which may occur independently of test expectancy. Weinstein et al.'s account, however, suggests that decreased expectancy may reduce attention in a more global way. While these accounts are related, they are not identical. However, Experiments 2 and 3 were not designed to test between these two accounts.

The reading times for Experiment 2 offer some support for the idea that JOLs influenced future learning behavior. In Experiment 1, reading times remained stable across sections for the tested conditions, but dropped following the first section for the non-tested condition. However, in Experiments 2 and 3, reading times remained stable across sections for *all* conditions. While it is unclear whether this lack of reduction in reading times for the non-tested conditions in Experiments 2 and 3 is the result of enhanced attention or mobilization of more effective (and thus more time-consuming) strategies, this provides some preliminary evidence that JOLs influenced subsequent study behavior.

It is also important to note that Yang et al. (2017) required aggregate judgments of learning after each study list in two experiments and did find a forward testing effect. However, like the present studies, they did report one experiment including JOLs (Experiment 3) and one without JOLs (Experiment 2), which both used the same materials. I examined the size of the forward testing effect between these two experiments in order to determine whether requiring JOLs reduced (but did not eliminate) the forward testing effect. The authors did not report standardized measures of effect size for the forward testing effect from either of these experiments, but the raw mean difference between their interim test group (most equivalent to prior cued-recall in the present studies) and the no interim test group (equivalent to no prior test) was higher when no judgments of learning were required (M_{Diff} = .28) than when they were provided (M_{Diff} = .16). However, this comparison should be interpreted with caution for two

reasons. First, these mean differences were obtained from separate experiments, so differences in the size of the forward testing effect could be due to sampling variability. Second, encoding time was self-paced when participants were not required to provide JOLs, but was fixed when they did provide them. This fixed encoding duration was quite short (4 seconds) relative to when participants could spend as long as they liked encoding the materials (approximately 10 seconds for the first list of materials).

This difference between studies leads to a separate explanation as to why Yang et al. (2017) observed a forward testing effect despite having participants provide JOLs whereas I did not. In Experiments 2 and 3 in this dissertation, encoding time was always self-paced. As discussed previously, reading times dropped across sections in the no prior test condition when JOLs were not given in Experiment 1, but remained stable when JOLs were provided. Thus, JOLs did have an impact on encoding behavior in Experiments 2 and 3. In contrast, because encoding time was fixed in Yang et al.'s Experiment 3, the JOLs could not enhance encoding duration in their no-interim test condition. Therefore, an important question for future research is how evaluative judgments influence performance when learning is self-regulated or experimenter paced.

One final difference between Yang et al.'s (2017) studies requiring JOLs and those reported in this dissertation is that I asked participants four JOL questions after each passage (16 total) whereas Yang et al only asked for one JOL after each list (four total). It could be that increasing the number of metacognitive judgments participants were required to make after each section also increased the likelihood of reactivity/increased test expectancy. If this is the case, it is unclear whether the effects of each judgment are additive (e.g., two judgments should be more likely to improve performance than one) or whether there is some sort of "critical mass" for these

judgments at which subsequent encoding is improved. Again, this is an interesting question for future research.

In addition to the lack of a forward testing effect in Experiments 1 and 2, another puzzling finding from the criterial test analyses is that the main effect of criterial test was absent. That is, the finding that performance was better on multiple-choice tests than cued-recall tests was not observed in either Experiments 2 or 3. Even more puzzling is that this null effect was only observed on the criterial test. For the tests for Sections 1-3, performance was always worse on multiple-choice tests than cued-recall tests. In order to determine if this null effect was the result of insufficient power, I conducted a combined analysis of Experiments 2 and 3 (collapsing across prior test type) between the criterial multiple-choice and criterial cued-recall test. A 2 (Criterial Test Type: Cued-Recall or Multiple-Choice) x 2 (Experiment 2 or 3) yielded no significant main effects nor an interaction, $F's < 2.17$. I buttressed this analysis with a Bayesian ANOVA examining the effect of prior review type and criterial test type on performance in Experiments 2 and 3 combined. There was substantial evidence in favor of the null hypothesis for the main effect of criterial test type, $BF_{01} = 9.54$[11]. Therefore, this difference does appear to reflect a real null difference between the test types. One potential hypothesis for why this occurred is related to the explanation for the lack of a forward testing effect. Specifically, while the JOLs seemed to particularly influence participants in the no-prior test condition, it is possible that asking participants in the tested conditions to provide JOLs could have improved performance in those conditions as well via the same mechanism (albeit to a lesser degree than those who did not take interpolated tests). However, this is purely speculative, and cannot explain why the same effect was not absent for the Section 3 test in Experiments 2 or 3.

[11] This analysis also provided strong evidence in support of the null hypothesis for the main effect of prior test type (i.e., the forward testing effect), BF_{01}=51.21.

Whereas there was no forward testing effect observed in Experiments 2 and 3, the JOLs provided evidence that testing does impact metacognitive judgments. First, in Experiment 2, there was evidence that participants were sensitive to the difficulty of the multiple-choice and cued-recall test formats. Specifically, in both Experiments 2 and 3, participants rated future performance higher when the JOL queried multiple-choice test performance than when it queried cued-recall performance. Second, in Experiment 2, there was evidence that prior testing (of either format) decreased JOLs across sections, particularly from the first to the second section. That is, while JOLs remained stable across sections when participants were not tested, those who received tests reduced their confidence in future performance.

Including the present dissertation, only three studies have examined how JOLs change in the context of a paradigm testing the forward testing effect. Yang et al. (2017) found *lower* JOLs for the non-tested rather than the tested conditions, and Szpunar et al. (2014) found no difference in JOLs between the non-tested and the tested conditions. Thus, the finding that JOLs were *higher* in the non-tested than tested conditions is a novel finding. This difference between the testing and non-testing conditions was also especially pronounced for predictions about future cued recall performance. However, there was a discrepancy between Experiments 2 and 3 regarding this analysis. In Experiment 3, JOLs declined across Sections regardless of condition and were higher for multiple-choice predictions than cued-recall, but there were no other main effects or significant interactions. In order to determine the effect of prior review type across sections with greater power, I combined Experiments 2 and 3 and re-ran the analysis (excluding the additional difficult multiple-choice condition from Experiment 3). This combined analysis largely replicated the results of Experiment 2 (in which prior testing reduced JOLs to a greater degree than no prior testing), so I will not discuss these results further. Therefore, the important

take-away from the JOLs is that they are sensitive to test format and tend to decrease across sections for the tested conditions.

One important contribution of Experiment 3 was to test whether increasing the difficulty of a multiple-choice test could impart forward testing benefits similar to cued-recall. In Experiments 1 and 2, the multiple-choice tests were much easier than the cued-recall tests. Under my interpretation of the metacognitive account, it is the *difficulty* of the interpolated tests, rather than the format *per se*, that should influence future performance. Thus, I predicted that difficult multiple-choice questions would enhance new learning to a greater degree than easy multiple-choice tests. Unfortunately, it appears that requiring participants to provide JOLs (as they did in Experiments 2 and 3) eliminated the forward effect of testing, making a test of this prediction difficult to interpret. It is important to note that accuracy was lower on the difficult multiple-choice tests for Sections 1-3 than for the easy multiple-choice tests (although the cued-recall tests still resulted in the lowest accuracy). Given that participants did find these tests more difficult, future research should examine the forward testing effect using these materials (without requiring JOLs).

Finally, the present experiments examined how interpolated testing would influence later test performance for both previously tested and previously untested items. Across all three experiments, there was a large benefit of retrieval practice, with higher performance on previously tested than previously untested items, and higher performance in the previously tested conditions than the previously non-tested condition. Regardless of whether a forward testing effect was present on the criterial tests (e.g., Experiment 1 relative to Experiments 2 and 3) there was no difference in performance between previously tested and previously untested items based on prior test condition. There was also no forward testing effect observed for *new* items from the

criterial section. That is, the benefit of prior testing did not seem to persist across a 15-minute delay. While results have been mixed regarding the persistence of the forward testing effect across a delay, this finding is somewhat unexpected. However, the only other study to investigate the forward testing effect (Wissman & Rawson, 2015) by using text passages (rather than word lists) demonstrated a reduction in the effect over a retention interval. Other studies that have found the effect to persist across a delay (e.g., Chan, Manley, Davis, & Szpunar, 2018) have used word lists. Whether the type of material encoded moderates the effect of delay is unclear, but it may be the case that the forward testing effect is more fragile over time for complex materials.

Taken together, the three studies conducted for this dissertation provide some preliminary evidence in favor of the metacognitive account for the forward effect of testing. As mentioned in the Introduction, the present studies were not designed to explicitly rule out any of the other extant theories for the forward effect of testing. My intent was to provide a direct test of the viability of the metacognitive account while simultaneously examining how multiple-choice tests might differentially affect new learning. Even so, support for the various theories under the proposed paradigm bears mentioning.

Under a proactive interference account (Szpunar et al., 2008), both recognition and recall should reduce proactive interference from prior sections/lists and should thus benefit new learning equally. Second, a contextual segregation account would make a similar prediction (Bauml & Kliegl, 2013; Lehman, Smith, & Karpicke, 2014; Pastotter & Bauml, 2014; Sahakyan & Hendricks, 2012). There is no a priori reason to assume that cued-recall tests would isolate study contexts better than multiple-choice tests, and this class of theories would predict an equal beneficial effect of either test. Again, this is not what occurred in Experiment 1, suggesting that

such a theoretical account is not sufficient on its own to encompass the mechanism behind the forward effect of testing. Similarly, I would argue that one might make the same null prediction for integration accounts. In general, one could argue that multiple-choice tests encourage integration as much as cued-recall tests, suggesting a null effect between the two types of testing. The current studies cannot test this possibility, but again, this theoretical approach seems unlikely to be able to account for the results on its own.

One important finding from the present studies is that requiring learners to predict their future performance completely eliminated the forward effect of testing. Importantly, this appeared to occur because of an increase in performance in the no-prior test condition. It is unclear how theories of proactive interference, context segregation, and integration would account for these data. Rather, this provides some support for the idea that testing enhances new learning by leading learners to adjust their subsequent encoding strategies. Whether this is by increasing the use of explicit strategy use (e.g., increasing the encoding of relational information, Chan, Manley, Davis, & Szpunar, 2018), or simply by decreasing inattention in a global manner (Szpunar et al., 2013) remains to be tested. It is also important to note that while evaluative judgments did affect criterial test performance, retrieval practice in general had the largest influence on the final cumulative test. That is, performance was highest for repeatedly tested questions, and in the tested conditions. Therefore, while evaluative judgments have short-term impacts on performance, retrieval of any type is essential to long-term learning.

The findings from the current studies have important applied, as well as theoretical, implications. For educators who administer clicker questions during learning (or students who practice during textbook reading) the results of Experiment 1 may suggest that in-class clicker questions should *not* be multiple-choice, and that instead learners and educators should

administer more difficult recall questions *regardless of the format* of the criterial or cumulative exam. Unfortunately, this is not always possible to do in large lecture classes, or when labor hours are limited. Of course, it would be advisable given the results of Experiment 1, that instructors always use cued-recall testing to quiz students. However, whereas prior multiple-choice testing did not enhance learning to the same degree as cued-recall testing in Experiment 1, there were benefits of prior multiple-choice testing relative to no prior testing. First, reading times remained stable (and equivalent to the prior cued-recall condition) across sections. Therefore, multiple-choice testing may aid students in paying attention during reading and during lecture classes. Second, multiple-choice testing did have an impact on JOLs, such that confidence in future performance was reduced after the first test opportunity. Given that overconfidence can lead students to make incorrect decisions about when or how to restudy learned material, this is important. Multiple-choice tests can improve calibration, which may have positive impacts down-stream from the test opportunity itself. The present studies did not include an additional restudy opportunity. However, in a real learning setting, students will be exposed and re-exposed to material as often as they choose restudy. If multiple-choice tests encourage students to restudy material at all or more often, it may still prove useful as a learning tool.

Concluding Remarks

To conclude, interpolated testing does appear to be an easy and cost-efficient way to enhance STEM learning in college classrooms, particularly when tests are more difficult. Given the call for more STEM graduates, coupled with reduced funding for public universities, the discovery and application of budget-conscious techniques to improve learning is paramount. In the present series of studies, cued-recall tests enhanced learning to a greater degree than

multiple-choice. But, to capture student attendance and engagement, less effortful tests (e.g., multiple-choice) may serve an important purpose—to increase attendance, guide study habits and metacognitive knowledge, and improve performance.

REFERENCES

Anderson, M.C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*(4), 415-445. doi:10.1016/j.jml.2003.08.006

Arkes, H.R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior & Human Decision Processes*, 39, 135–144.

Bauml, K.H.T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language, 68*(1), 39-53. doi:10.1016/j.jml.2012.07.006

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

Butler, A.C., & Roediger, H.L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. doi: 10.3758/MC.36.3.604

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. doi: 10.3758/BF03193405

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*(2), 153-170. doi:10.1016/j.jml.2009.04.004

Chan, J.C.K., & McDermott, K.B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 431-437. doi: 10.1037/0278-7393.33.2.431

Chan, J.C.K., McDermott, K.B., & Roediger, H.L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553-571. doi:10.1037/0096-3445.135.4.553

Chan, J.C.K., Meissner C.A., & Davis, S.D. (2018) Test-Potentiated New Learning: A Meta-Analytic Review. *Psychological Bulletin.*

Chan, J.C.K., Manley, K.D., Davis, S.D., & Szpunar, K.K., (2018) Testing Potentiates New Learning Across a Retention Interval and a Lag: A Strategy Change Perspective. *Journal of Memory and Language.*

Chan, J.C.K., McDermott, K.B., & Roediger, H.L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553-571. doi:10.1037/0096-3445.135.4.553

Chen, X., and Soldner, M. (2013) STEM Attrition: College Students' Into and Out of STEM Fields. US Department of Education.

Davis, S.D., & Chan, J.C.K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition,* doi: 10.1037/xlm0000126

Dougherty, M.R., Robey, A.M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory & Cognition, 46*, 558-565. doi:10.3758/s13421-018-0791-y

Dougherty, M.R., Scheck, P., Nelson, T.O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*(6), 1096-1115. doi:10.3758/BF03193216

Dunlosky, J., & Rawson, K.A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271-280. doi: /10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., & Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:10.1177/1529100612453266

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513-541. doi: 10.1016/0749-596X(91)90025

Jacoby, L.L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*(3), 306-340. doi: 10.1037/0096-3445.110.3.306

Jing, H.G., Szpunar, K.K., & Schacter, D.L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*, 305–318.

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition: An International Journal, 19*(1), 251-264. doi:10.1016/j.concog.2009.12.010

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

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609-622. doi:10.1037/0278-7393.32.3.609

Kouyoumdjian, H. (2004). Influence of unannounced quizzes and cumulative exam on attendance and study behavior. *Teaching of Psychology, 31*(2), 110-111.

Lehman, M., Smith, M.A., & Karpicke, J.D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1787-1794. doi:10.1037/xlm0000012

Little, J.L., Bjork, E.L., Bjork, R.A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337-1344. doi:10.1177/0956797612443370

Mandler, G. (2008). Familiarity breeds attempts: A critical review of dual-process theories of recognition. *Perspectives on Psychological Science, 3*(5), 390-399. doi:10.1111/j.1745-6924.2008.00087.x

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*(2), 196-204. doi:10.3758/BF03197095

McDermott, K.B., Agarwal, P.K., D'Antonio, L., Roediger., H.L., III, & McDaniel, M.A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*(1), 3-21. doi:.1037/xap0000004

Mitchum, A.L., Kelley, C.M., & Fox, M.C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200-219. doi:10.1037/a0039923

National Center for Education Statistics. (n.d.).

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Pastotter, B., & Bauml, K.T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*, 5.

Pastotter, B., Schicker, S., Niedernhuber, J., & Bauml, K.H.T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37,* 287–297. doi:10.1037/a0021801

Pastotter, B., Weber, J., Bauml, K.H.T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*(2), 280-285. doi:10.1037/a0031797

Rhodes, M.G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky, & S. K. Tauber (Eds.), *The oxford handbook of metamemory; the oxford handbook of metamemory* (pp. 65-80, Chapter xv, 574 Pages) Oxford University Press, New York, NY.

Roediger, H.L., III, & Marsh, E.J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155-1159. doi: 10.1037/0278-7393.31.5.1155

Roediger, H.L., III, & Karpicke, J.D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Rowland, C.A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:10.1037/a0037559

Sahakyan, L., Delaney, P.F., & Kelley, C.M. (2004). Self-evaluation as a moderating factor of strategy change in directed forgetting benefits. *Psychonomic Bulletin & Review, 11*(1), 131-136. doi: 10.3758/BF03206472

Sahakyan, L., & Hendricks, H.E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition, 40*(6), 844-860. doi:10.3758/s13421-012-0198-0

Saunders, J., & MacLeod, M.D. (2002). New evidence on the suggestibility of memory: The role of retrieval-induced forgetting in misinformation effects. *Journal of Experimental Psychology: Applied, 8*(2), 127-142. doi:10.1037/1076-898X.8.2.127

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

Soderstrom, N.C., Clark, C.T., Halamish, V., & Bjork, E.L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553-558. doi:10.1037/a0038388

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment and Evaluation in Higher Education, 30*, 331-347.

Szpunar, K.K., Jing, H.G., & Schacter, D.L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition*, *3*, 161–164.

Szpunar, K.K., Khan, N.Y., & Schacter, D.L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 6313–6317.

Szpunar, K.K., McDermott, K.B., & Roediger, H.L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392-1399. doi:10.1037/a0013082

Thiede, K.W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*(2), 290-302. doi:10.1037/0022-0663.86.2.290

Wais, P.E., Mickes, L., & Wixted, J.T. (2008). Remember/know judgments probe degrees of recollection. *Journal of Cognitive Neuroscience, 20*(3), 400-405. doi: 10.1162/jocn.2008.20041

Weinstein, Y., Gilmore, A.W., Szpunar, K.K., & McDermott, K.B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1039-1048. doi:10.1037/a0036164

Weinstein, Y., McDermott, K.B., & Szpunar, K.K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review, 18*(3), 518-523. doi:10.3758/s13423-011-0085-x

Wilder, D.A., Flood, W.A., & Stromsnes, W. (2001). The use of random extra credit quizzes to increase student attendance. *Journal of Instructional Psychology, 28*(2), 117-120.

Wissman, K.T., & Rawson, K.A. (2015). Grain size of recall practice for lengthy text material: fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 439–455.

Wissman, K.T., Rawson, K.A., & Pyc, M.A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*, 1140–1147

Wixted, J.T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. Psychonomic Bulletin & Review, 11(4), 616-641. doi: 10.3758/BF03196616

World Economic Outlook Database, October 2016. https://ww.imf.org/external/pubs/ft/weo/2016/weodata.index.aspx

Yang, C., Potts, R., & Shanks, D.R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*(3), 263-277. doi: 10.1037/xap0000122

Zeidener, M. (1998). Test Anxiety: The State of the Art. Springer Science.

APPENDIX A. TEXT PASSAGES

History of the Laser and Laser Basics (Word Count: 753)

We can trace the birth of lasers right back to the first two decades of the 20th century. That's when Albert Einstein figured out the quantum theory of light and photons (in 1905) and the mechanism of stimulated emission (in 1917)—the two key components of laser science. But it was another four decades before the first practical laser actually appeared.

Lasers evolved from masers, which are similar but produce microwaves and radio waves instead of visible light. Masers were invented in the 1950s by Charles Townes and Arthur Schawlow, both of whom went on to win the Nobel Prize in Physics for their work.

But did they invent the laser? In 1957, one of Townes' graduate students, Gordon Gould, sketched in his lab notebook an idea for how a visible light version of the maser could work, coining the word "laser" that we've used ever since. Unfortunately, he didn't patent his idea at the time and had to devote the next 20 years of his life to legal battles, eventually gaining a patent for part of the laser invention and substantial back royalties in 1977.

Although Townes and Schawlow are often credited with inventing lasers, the first person to *build* a working, visible light laser was actually Theodore Maiman, who has never really gained the recognition he deserved: his original writeup of his work was rejected by the journal Physical Review Letters and, despite twice being nominated for the Nobel Physics Prize, he never won the ultimate accolade.

Lasers are amazing light beams powerful enough to zoom miles into the sky or cut through lumps of metal. But early inventors of the laser had no idea what to do with lasers; famously, they were described as "a solution looking for a problem." Today, we all have lasers in our homes (in CD and DVD players), in our offices (in laser printers), and in the stores where we shop (in barcode scanners). Our clothes are cut with lasers, we fix our eyesight with them, and we send and receive emails over the Internet with signals that lasers fire down fiber-optic cables. Whether we realize it or not, all of us use lasers all day long, but how many of us really understand what they are or how they work?

Lasers are more than just powerful flashlights. The difference between ordinary light and laser light is like the difference between ripples in your bathtub and huge waves on the sea. You've probably noticed that if you move your hands back and forth in the bathtub you can make quite strong waves. If you keep moving your hands in step with the waves you make, the waves get bigger and bigger. Imagine doing this a few million times in the open ocean. Before long, you'd have mountainous waves towering over your head! A laser does something similar with light waves. It starts off with weak light and keeps adding more and more energy so the light waves become ever more concentrated.

If you've even seen a laser in a science lab, you'll have noticed two very important differences straightaway. First, where a flashlight produces "white" light (a mixture of all different colors, made by light waves of all different frequencies), a laser makes what's called monochromatic

light (of a single, very precise frequency and color—often bright red or green or an invisible "color" such as infrared or ultraviolet). Second, where a flashlight beam spreads out through a lens into a short and fairly fuzzy cone, a laser shoots a much tighter, narrower beam over a much longer distance (we say it's highly collimated). There's a third important difference you won't have noticed: Where the light waves in a flashlight beam are all jumbled up, (with the crests of some beams mixed with the troughs of others), the waves in laser light are exactly in step: the crest of every wave is lined up with the crest of every other wave. We say laser light is coherent. Think of a flashlight beam as a crowd of commuters, pushing and shoving, jostling their way down the platform of a railroad station; by comparison, a laser beam is like a parade of soldiers all marching precisely in step. These three things make lasers precise, powerful, and amazingly useful beams of energy.

Essentially, a laser is effectively a machine that makes billions of atoms pump out trillions of photons (light particles) all at once so they line up to form a really concentrated light beam.

Making a Laser Beam (Word Count: 754)

To make a laser, we need two basic parts. First, we need a load of atoms (a solid, liquid, or gas) with electrons in them that we can stimulate. This is known as the medium or, sometimes, the amplifying or "gain" medium (because gain is another word for amplification). Second, we need something to stimulate the atoms with, such as a flash tube (like the xenon flash lamp in a camera) or another laser. A typical red laser would contain a long crystal made of ruby with a flash tube wrapped around it. The flash tube looks a bit like a fluorescent strip light, only it's coiled around the ruby crystal and it flashes every so often like a camera's flash lamp.

To create a beam of laser light, first a high-voltage electric supply makes the tube flash on and off. Every time the tube flashes, it "pumps" energy into the ruby crystal. The flashes it makes inject energy into the crystal in the form of photons. Atoms in the ruby crystal soak up this energy in a process called absorption. Atoms absorb energy when their electrons jump to a higher energy level. After a few milliseconds, the electrons return to their original energy level (ground state) by giving off a photon of light. This is called spontaneous emission.

Next, the photons that atoms give off zoom up and down inside the ruby crystal, traveling at the speed of light. Every so often, one of these photons stimulates an already excited atom. When this happens, the excited atom gives off a photon and we get our original photon back as well. This is called stimulated emission. Now one photon of light has produced two, so the light has been amplified (increased in strength). In other words, "light amplification"(an increase in the amount of light) has been caused by "stimulated emission of radiation" (hence the name "laser", because that's exactly how a laser works!) A mirror at one end of the laser tube keeps the photons bouncing back and forth inside the crystal. A partial mirror at the other end of the tube bounces some photons back into the crystal but lets some escape. The escaping photons form a very concentrated beam of powerful laser light.

Lasers make electromagnetic radiation, just like ordinary light, radio waves, X rays, and infrared. Although it's still produced by atoms, they make ("emit") it in a totally different way, when

electrons jump up and down inside them. We can think of electrons in atoms sitting on energy levels, which are a bit like rungs on a ladder. Normally, electrons sit at the lowest possible level, which is called the atom's ground state. If you fire in just the right amount of energy, you can shift an electron up a level, onto the next rung of the "ladder." That's called absorption and, in its new state, we say the atom's excited—but it's also unstable. It very quickly returns to the ground state by giving off the energy it absorbed as a photon (a particle of light). That's why we call this process spontaneous emission of radiation: the atom is giving off light (emitting radiation) all by itself (spontaneously).

Normally, a typical bunch of atoms would have more electrons in their ground states than their excited states, which is one reason why atoms don't spontaneously give off light. But when we excite the atoms, this places their electrons in excited states. Now suppose also that we could maintain our atoms in this state for a little while so they didn't automatically jump back down to their ground state (a temporarily excited condition known as a meta-stable state). Then we'd find something really interesting. If we fired a photon with just the right energy through our bunch of atoms, we'd cause one of the excited electrons to jump back down to its ground state, giving off both the photon we fired in and the photon produced by the electron's change of state. Because we're stimulating atoms to get radiation out of them, this process is called stimulated emission. We get two photons out after putting one photon in, effectively doubling our light and amplifying it (increasing it). These two photons can stimulate other atoms to give off more photons, so, pretty soon, we get a cascade of photons—a chain reaction—throwing out a brilliant beam of pure, coherent laser light. What we've done here is amplify light using stimulated emission of radiation—and that's how a laser gets its name.

Laser Light, Types of Lasers, and Safety Classifications (Word Count: 748)

Why do lasers make a single color and a coherent beam? It boils down to the idea that energy can only exist in fixed packets, each of which is called a quantum. It's a bit like money. You can only have money in multiples of the most basic unit of your currency, which might be a cent, penny, rupee, or whatever. You can't have a tenth of a cent or a twentieth of a rupee, but you can have 10 cents or 20 rupees. The same is true of energy, and it's particularly noticeable inside atoms.

Like the rungs on a ladder, the energy levels in atoms are in fixed places, with gaps in between them. You can't put your foot anywhere on a ladder, only on the rungs; and in exactly the same way, you can only move electrons in atoms between the fixed energy levels. To make an electron jump from a lower to a higher level, you have to feed in a precise amount (quantum) of energy, equal to the difference between the two energy levels. When electrons flip back down from their excited to their ground state, they give out the same, precise amount of energy, which takes the form of a photon of light of a particular color. Stimulated emission in lasers makes electrons produce a cascade of identical photons—identical in energy, frequency, wavelength—and that's why laser light is monochromatic. The photons produced are equivalent to waves of light whose crests and troughs line up (in other words, they are "in phase")—and that's what makes laser light coherent.

Since we can excite many different kinds of atoms in many different ways, we can (theoretically) make many different kinds of lasers. In practice, there are only a handful of common kinds, of which the five best known are solid-state, gas, liquid dye, semiconductor, and fiber.

Solids, liquids, and gases are the three main states of matter—and give us three different kinds of lasers. Solid-state lasers are like the ones I illustrated up above. The medium is something like a ruby rod or other solid crystalline material, and a flashtube wrapped around it pumps its atoms full of energy. To work effectively, the solid has to be doped, a process that replaces some of the solid's atoms with ions of impurities, giving it just the right energy levels to produce laser light of a certain, precise frequency. Solid-state lasers produce high-powered beams, typically in very brief pulses.

Gas lasers, by contrast, produce continuous bright beams using compounds of noble gases (in what are called excimer lasers) or carbon dioxide (CO_2) as their medium, pumped by electricity. CO_2 lasers are powerful, efficient, and typically used in industrial cutting and welding.

Liquid dye lasers use a solution of organic dye molecules as the medium, pumped by something like an arc lamp, a flash lamp, or another laser. Their big advantage is that they can be used to produce a broader band of light frequencies than solid-state and gas lasers, and they can even be "tuned" to produce different frequencies.

While solid, liquid, and gas lasers tend to be large, powerful, and expensive, semiconductor lasers are cheap, tiny, chip-like devices used in things like CD players, laser printers, and barcode scanners. They work like a cross between a conventional Light-emitting diode (LED) and a traditional laser. Like an LED, they make light when electrons and "holes" (effectively, "missing electrons") hop about and join together; like a laser, they generate coherent, monochromatic light. That's why they're sometimes referred to as laser diodes (or diode lasers). You can read more about them in our separate article about semiconductor laser diodes.

Finally, fiber lasers work their magic inside optical fibers; in effect, a doped fiber-optic cable becomes the amplifying medium. They're powerful, efficient, reliable, and make it easy to pipe laser light to wherever it's needed.

Lasers are also grouped into seven classes depending on the potential for the beam to cause harm. The hazard and hence the classification depends on the wavelength, power, energy and pulse characteristics. The class of the laser can be used to help decide what safety control measures are required when using the laser. The Accessible Emission Limit (AEL) is the maximum level of laser radiation which a laser can emit (and be accessible) at any time after its manufacture. The AEL depends on the wavelength, exposure duration and the viewing conditions and specifies the maximum output within each laser class.

Laser Uses (Word Count: 750)

When Theodore Maiman developed the first practical laser, few people realized how important these machines would eventually become. *Goldfinger*, the 1964 James Bond movie, offered a

tantalizing glimpse of a future where industrial lasers could slice like magic through anything in their path—even secret agents! Later the same year, reporting on the award of the Nobel Prize in Physics to the laser pioneer Charles Townes, *The New York Times* suggested that "a laser beam could, for example, carry all the radio and television programs in the world plus several hundred thousand telephone calls simultaneously. It is used extensively for range-finding and missile-tracking." Over half a century later, applications like this—precision tools, digital communication, and defense—remain among the most important uses of lasers.

Cutting tools based on CO_2 lasers are widely used in industry: they're precise, easy-to-automate, and, unlike knives, never need sharpening. Where pieces of cloth were once cut by hand to make things like denim jeans, now fabrics are chopped by robot-guided lasers. They're faster and more accurate than humans and can cut multiple thicknesses of fabric at once, which improves efficiency and productivity. The same precision is equally important in medicine: doctors routinely use lasers on their patients' bodies for everything from blasting cancer tumors and cauterizing blood vessels to correcting problems with people's vision.

Laser eye surgery itself is relatively simple and it's easy to understand. First, your eye is treated with an anesthetic, then your eyelids are gripped and held open by a suction frame. The same device pulls on the cornea and holds it securely in place ready for the surgery. It's not exactly painful, though it's not comfortable either. Next, using tiny bursts of powerful light lasting no more than a few nanoseconds, the laser cuts a flap in your cornea. (In some procedures, the flap is cut by a microscopic knife called a microkeratome blade.) Think of the flap as a bit like a small door in the surface of the cornea: it's cut on three sides but the fourth side is left attached to form a kind of hinge. The flap is then lifted up and folded back on the hinge to expose the inner corneal tissue underneath. Under computer control, the laser then reshapes the cornea under the flap. Once that's done, the flap is replaced and it will slowly reattach itself without any need for stitches or surgical adhesives.

Lasers also form the bedrock of all kinds of 21st-century digital technology. Every time you swipe your shopping through a grocery store barcode scanner, you're using a laser to convert a printed barcode into a number that the checkout computer can understand. When you watch a DVD or listen to a CD, a semiconductor laser beam bounces off the spinning disc to convert its printed pattern of data into numbers; a computer chip converts these numbers into movies, music, and sound. Along with fiber-optic cables, lasers are widely used in a technology called photonics—using photons of light to communicate, for example, to send vast streams of data back and forth over the Internet.

The military has long been one of the biggest users of this technology, mainly in laser-guided weapons and missiles. Despite its popularization in movies and on TV, the sci-fi idea of laser weapons that can cut, kill, or blind an enemy remained fanciful until the mid-1980s. In 1981, *The New York Times* went so far as to quote one "military laser expert" saying: "It's just silly. It takes more energy to kill a single man with a laser than to destroy a missile." Two years later, long-range laser weapons officially became the bedrock of US President Ronald Reagan's controversial Strategic Defense Initiative (SDI), better known as the "Star Wars program". The original idea was to use space-based, X ray lasers (among other technologies) to destroy

incoming enemy missiles before they had time to do damage, though the plan gradually fizzled out following the collapse of the Soviet Union and the end of the Cold War. Even so, defense scientists have continued to transform laser-based missiles from science fiction into reality. In 2014, the US Navy successfully tested LaWS (Laser Weapon System) onboard a ship in the Persian Gulf. Using solid-state lasers pumped by LEDs, it's designed to damage or destroy enemy equipment more cheaply and precisely than conventional missiles, and expected to be rolled out more widely from 2016 onward. Meanwhile, the development of space lasers continues, though none have so far been deployed.

APPENDIX B. TEST QUESTIONS AND ANSWERS

Section 1

	Question	Answer
1	What did lasers evolve from?	Masers
2	What type of particles do masers emit?	Microwaves and radio waves
3	Who invented the laser?	Gordon Gould
4	Why did Gordon Gould have to fight so many legal battles?	He didn't patent the laser
5	Who is Theodore Maiman?	He built the first laser
6	What is the name for light of a very precise wavelength and color?	Monochromatic
7	A "highly collimated" beam of light is very_____	Narrow
8	_____ light is when different wavelengths of light are lined up precisely.	Coherent

Section 2

1	What part of the laser emits light to begin the process of creating a laser beam?	Flash tube
2	What part of the laser is stimulated by the flash tube?	Gain medium
3	What would a red laser's gain medium be made from?	Ruby
4	What happens when the flash tube goes off?	Energy is pumped into the crystal
5	What is the normal state of an electron?	Ground state
6	When electrons give off a photon of light, this is called _____.	Spontaneous emission
7	What does the L.A. in laser stand for?	Light amplification
8	What does the R in laser stand for?	Radiation

Section 3

1	What is a quantum?	A fixed packet of energy.
2	What is a solid-state laser?	A laser where the gain medium is a solid
3	What is a gas laser?	A laser whose gain medium is a gas
4	What are CO2 lasers used for?	Welding, industrial cutting
5	What is the advantage of using liquid dye lasers?	They produce a broader spectrum of light
6	What is another name for semiconductor lasers?	Laser diodes
7	How many safety classes of laser are there?	Seven
8	What is the Accessible Emission Limit (AEL)?	The maximum energy a laser can emit

Section 4

1	Who won the Nobel Physics prize for developing the laser?	Charles Towne
2	What kind of lasers are used for cutting materials?	CO2 or gas lasers
3	What part of the eye is cut in laser eye surgery?	The cornea

4	What is a microkeratome blade?	A very small knife.
5	What kinds of 21st century technology use lasers?	DVD's and barcode scanners
6	What is photonics?	Using photons of light to communicate
7	What industry is the biggest user of lasers?	The military
8	How many space lasers have been deployed?	None

APPENDIX C. IRB APPROVAL

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 9/2/2016

To: Dr. Jason Chun Kit Chan
W112 Lagomarcino Hall

From: Office for Responsible Research

Title: Tricks to Improve Your Memory

IRB ID: 15-609

Approval Date:	8/30/2016	**Date for Continuing Review:**	1/24/2018
Submission Type:	Modification	**Review Type:**	Expedited

The project referenced above has received approval from the Institutional Review Board (IRB) at Iowa State University according to the dates shown above. Please refer to the IRB ID number shown above in all correspondence regarding this study.

To ensure compliance with federal regulations (45 CFR 46 & 21 CFR 56), please be sure to:

- **Use only the approved study materials** in your research, **including the recruitment materials and informed consent documents that have the IRB approval stamp.**
- **Retain signed informed consent documents for 3 years after the close of the study,** when documented consent is required.
- **Obtain IRB approval prior to implementing** any **changes** to the study by submitting a Modification Form for Non-Exempt Research or Amendment for Personnel Changes form, as necessary.
- **Immediately inform the IRB of (1) all serious and/or unexpected adverse experiences** involving risks to subjects or others; and (2) **any other unanticipated problems involving risks** to subjects or others.
- **Stop all research activity if IRB approval lapses,** unless continuation is necessary to prevent harm to research participants. Research activity can resume once IRB approval is reestablished.
- **Complete a new continuing review form** at least three to four weeks prior to the **date for continuing review** as noted above to provide sufficient time for the IRB to review and approve continuation of the study. We will send a courtesy reminder as this date approaches.

Please be aware that IRB approval means that you have met the requirements of federal regulations and ISU policies governing human subjects research. **Approval from other entities may also be needed.** For example, access to data from private records (e.g. student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **IRB approval in no way implies or guarantees that permission from these other entities will be granted.**

Upon completion of the project, please submit a Project Closure Form to the Office for Responsible Research, 202 Kingland, to officially close the project.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.