THE TESTING EFFECT UNDER DIVIDED ATTENTION: EDUCATIONAL APPLICATION

Zachary L. Buchin

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology and Neuroscience (Cognitive Psychology)

Chapel Hill
2018

Approved by:

Neil W. Mulligan

Jessica R. Cohen

Joseph B. Hopfinger

ABSTRACT

Zachary L. Buchin: The Testing Effect Under Divided Attention: Educational Application
(Under the direction of Neil W. Mulligan)

Taking a test enhances retention, often to a greater degree than restudying (i.e. the testing effect). Understanding how these encoding effects of retrieval differ from other forms of encoding is important when convincing educators of the benefits of testing. One potential difference relates to attention: Dividing attention disrupts memory encoding but typically has much less impact on retrieval. Less is known about the relative attentional demands of the encoding effects of retrieval. In three experiments, participants studied foreign language word pairs (Experiment's 1 and 2) or educational texts (Experiment 3), restudied or retrieved those materials under full attention (FA) or divided attention (DA), and then took a cued-recall test. A testing effect was found under FA and DA and the level of DA disruption was similar for both learning conditions. Consequently, the encoding effects of retrieval and restudy appear to be similarly susceptible to distraction when learning complex educational information.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

FA	Full Attention
DA	Divided Attention

CHAPTER 1: INTRODUCTION

Although the idea of both administering and taking frequent tests in the classroom may raise thoughts of student anxiety and lost time for new instruction, research has shown that tests not only assess but also enhance learning (e.g. Roediger & Karpicke, 2006). This widely documented finding, known as the testing effect, is revealed when retrieving information on a test improves later memory compared to simply restudying the same material (Chan & McDermott, 2007; Roediger & Butler, 2011; Roediger & Karpicke, 2006). The memory enhancement produced by testing has been observed in various situations, including lab studies, skill learning, multi-media stimuli tests, and classroom settings (Johnson & Mayer, 2009; Kromann, Jensen, & Ringsted, 2009; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Butler, 2011; Roediger & Karpicke, 2006).

Researchers exploring the testing effect typically use a three-phase paradigm to compare a restudy to a retrieval condition (e.g. Reodiger & Butler, 2011). In the first phase, participants are instructed to study material (e.g. words, word pairs, educational text) for a later memory test. In the critical second phase of the paradigm, participants either restudy the material or take a practice memory test. In the third phase, a final test is administered to compare memory for material restudied or retrieved during phase two. As discussed above, the material retrieved during phase two is generally better remembered than material restudied (i.e. the testing effect).

In an early study, Carrier and Pashler (1992) utilized this typical paradigm. In phase 1, participants studied nonsense-word/number pairs (experiment 1) or foreign language word

pairs (experiments 2 – 4). In phase 2, participants restudied some of the cue-target pairs and practiced retrieving the targets when shown the cue for other pairs. Results from a final cued-recall test (phase 3) demonstrated an advantage for items retrieved during phase 2 compared to items restudied during phase 2. Critically, this retrieval advantage was found for both nonsense-word/number pairs as well as more externally valid, foreign language word pairs. Overall, these results, along with other similar findings, demonstrate that testing a person's memory enhances later recall relative to restudying the same material (e.g. Carpenter, Pashler, Wixted, & Vul, 2008; Cull, 2000; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Zaromb & Roediger, 2010).

Despite the typical use of tests as measurement tools of memory, it is clear that retrieval does more than simply reveal the contents of memory; it also modifies memory representations (Buchin & Mulligan, 2017; Bjork & Bjork, 1992; Mulligan & Picklesimer, 2016; Roediger, & Butler, 2011; Roediger & Karpicke, 2006). The memory enhancement caused by retrieval (produced in the second phase of the typical paradigm and revealed on the final memory test) shows that retrieval has encoding (or perhaps, re-encoding) effects. Unlike the general testing effect itself, the mechanisms behind these retrieval-induced encoding effects have not been as widely studied. Determining exactly how the mnemonic benefits of retrieval operate, as well as their similarities to typical encoding processes (i.e. those that operate during periods of study or restudy), could be an important factor when persuading educators and students of the power of repeated testing.

However, it is helpful to first discuss accounts of the testing effect that offer predictions in regard to how similar the encoding effects of retrieval are to other encoding processes. For example, one theory argues that retrieval benefits subsequent memory because

it enhances elaborative processing (Carpenter, 2009, 2011). According to this account, when retrieving a target memory, participants elaborate on the memory trace by activating semantic associates of the cue and to-be-retrieved target. These elaborations then serve as additional retrieval routes for accessing the target memory on later memory tests (Carpenter & Yeung, 2017; Rawson, Vaughn, & Carpenter, 2015).

A second, somewhat similar account, argues that the benefits of retrieval (as compared to encoding) are due to more effortful processing of the target stimulus (e.g., Endres & Renkl, 2015; Kang, McDermott & Roediger, 2007; McDaniel, Roediger & McDermott, 2007; Pyc & Rawson, 2009; Stenlund, Sundstrom & Jonsson, 2016; also see Roediger & Butler, 2011; van den Broek et al., 2016). In other words, it is the increased mental effort brought on by memory retrieval that drives the testing effect (e.g. Endres & Renkel, 2015). For example, Pyc and Rawson (2009) found that increasing the number of items between an item and its practice trial (i.e. inter-stimulus interval) resulted in higher levels of later test performance. Additionally, requiring more successful retrieval practice (i.e. higher criterion levels) resulted in lower levels of later test performance (Pyc & Rawson, 2009).

If the memory benefit produced by testing is caused by greater elaboration or effort, then conditions that disrupt these processes should decrease the size of the testing effect. Attention offers an interesting domain in which to assess these accounts because of its differing effects on encoding and retrieval. Dividing attention during encoding usually produces substantial negative effects on memory, whereas the effects of dividing attention during retrieval are typically quite modest (Anderson, Craik, & Naveh-Benjamin, 1998; Craik, Govoni, Naveh-Benjamin, & Anderson, 1996; LeCompte, Neely, & Wilson, 1997;

Murdock, 1965; see Mulligan, 2008, for review). Before discussing the possible impact of attention on the testing effect, I will first review relevant empirical results concerning attention and general memory processes.

When assessing the effects of attention on memory in laboratory studies, researchers typically compare a divided attention (DA) condition, in which participants perform a memory task while also carrying out a secondary distractor task, to a full attention (FA) condition, in which only the memory task is performed (e.g. Craik et al., 1996; Lozito & Mulligan, 2010). As noted above, when attention is divided during encoding, later memory accuracy is greatly reduced (e.g. Anderson et al., 1998). However, if attention is divided during the memory test, there is typically little or no decrement in accuracy (e.g. Craik et al., 1996). Additionally, even when DA during retrieval does produce a significant reduction in test performance (e.g., Fernandes & Moscovitch, 2000; Hicks & Marsh, 2000; Mulligan & Lozito, 2006), the reduction is consistently smaller than the reduction found from DA during encoding (e.g. Anderson et al., 1998; Craik et al., 1996; Naveh-Benjamin, Craik, Perretta, & Tonev, 2000).

Critically, although engaging in a secondary task during retrieval produces little effect on memory accuracy, performance on the secondary task itself is impaired by retrieval and to a greater degree than when the secondary task is paired with encoding (e.g. Anderson et al., 1998; Craik et al., 1996; Naveh-Benjamin et al., 2000). For example, Craik and colleagues (1996) compared performance on a continuous reaction-time (CRT) task when performed alone to performance when performed alongside a memory task. Although the authors found that CRT performance decreased when paired with either encoding or retrieval, the decrement was significantly larger when paired with a retrieval task (e.g. Craik et al., 1996).

Therefore, Craik et al. (1996) concluded that although retrieval is resilient to distraction, it is not automatic, because it exacts attentional costs as shown by the reduced performance on the concurrent CRT. Consequently, Craik et al. (1996) characterized retrieval as attention-demanding but obligatory, arguing that retrieval takes precedence over other ongoing cognitive operations (e.g., the operations required by the concurrent task like the CRT).

Because of DA's differential effects on encoding and retrieval, it would be informative to examine the effects of DA on the encoding effects of retrieval. If the encoding effects of retrieval are similar to typical encoding, then attentional manipulations should reduce the benefit of retrieval practice on later memory. That is, dividing attention during the critical second phase of the standard testing effect paradigm should substantially reduce final memory recall similarly to DA during general encoding, or perhaps, to an even greater degree. This latter possibility is consistent with the predictions of the elaborative and effortful accounts of the testing effect discussed above. Research not only shows that DA generally disrupts memory encoding, but is especially disruptive of elaborative processing (Craik & Broadbent, 1983; Craik et al., 1996; Hasher & Zacks, 1979; McDowd & Craik, 1988; Mulligan, 2008). Likewise, testing effect accounts that focus on effort predict a similar net decrease, due to the importance of attention when exerting mental effort (Craik et al., 1996; Mulligan, 2008). Therefore, if testing effects are driven by greater elaboration or effort in the retrieval than restudy condition, DA should produce greater impairments to the encoding consequences of retrieval than restudy, yielding a net decrease in the size of the testing effect. Overall, views of the testing effect hinging on elaboration or effort predict a smaller testing effect under DA than under FA.

Alternatively, it may be that the encoding effects of retrieval are more similar to retrieval success itself. As noted above, DA during retrieval has much less impact on memory performance than DA during encoding, raising the possibility that the subsequent mnemonic effects of retrieval are also largely protected from the effects of distraction relative to a restudy condition. This prediction seems consistent with the idea that retrieval is attention-demanding but obligatory, and that it takes precedence over other ongoing tasks (e.g., in dual-task situations) whereas encoding typically does not (Anderson et al., 1998; Craik et al., 1996; Naveh-Benjamin et al., 2000). It is possible that one of the benefits of retrieval-based learning is that it is relatively protected from distraction compared to restudy (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). This possibility implies the opposite pattern of results than do the elaborative and effortful accounts of the testing effect. If the encoding effects of retrieval are resilient in the face of distraction, the size of the testing effect should increase under divided attention, as the restudy condition is more greatly harmed by distraction than the retrieval condition.

Until recently, there has been limited and conflicting research pertaining to DA and its effect on the mnemonic benefits of retrieval. Dudukovic, DuBrow, and Wagner (2009) had participants study a list of pictures (phase 1), followed by a recognition test for some of the old items, presented under FA or DA (phase 2). Accuracy on a final recognition test (phase 3) revealed a benefit of pictures tested during phase 2 (i.e. the testing effect). However, this benefit was reduced in the DA condition, compared to the FA condition. Overall, this led the authors to conclude that the encoding effects of retrieval are reduced under DA, similar to other encoding processes (Dudukovic et al., 2009; see also Dudukovic, Gottshall, Cavanaugh, & Moody, 2015).

A similar study was conducted to test if dividing attention during retrieval would act as a desirable difficulty thereby increasing later memory performance (Gaspelin, Ruthruff, & Pashler, 2013). First, participants studied Swahili-English word pairs (phase 1) and then took a cued-recall test (phase 2) under either FA or DA. Two days later, a final cued-recall test (phase 3) demonstrated that both attention conditions exhibited the same amount of forgetting. This finding led the authors to conclude that although DA does not impair the encoding effects of retrieval, it also does not act as a beneficial, desirable difficulty (Gaspelin et al., 2013).

Kessler et al. (2014) found a third conflicting result when assessing the effects of divided attention on retrieval. Participants studied pictures (phase 1) and then took a recognition memory test (under FA or DA) on half of the previously studied pictures (phase 2). Based on the results of a final recognition test (phase 3), the authors argued that DA during retrieval actually enhanced later memory for those pictures (Kessler et al., 2014).

Overall, the above experiments variously suggest that the encoding effects of retrieval are diminished, unaffected, or benefited by DA. However, none of the studies actually measured the testing effect, which requires a comparison of a restudy and a retrieval condition (e.g. Carpenter et al., 2008; Carrier & Pashler, 1992; Cull, 2000; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Zaromb & Roediger, 2010). Therefore, the results do not allow us to determine if the encoding effects of retrieval and restudy actually differ in terms of their reliance on attention. For example, because Gaspelin et al. (2013) examined potential desirable difficulties stemming from DA during retrieval, the authors did not include a restudy or untested control condition. Additionally, both Dudukovic et al. (2009)

and Kessler et al. (2014) used untested study items, rather than items restudied during phase 2, as their control condition.

Furthermore, Dudukovic et al. (2009) and Kessler et al. (2014) employed recognition tests during phase 2 retrieval, posing additional problems. First, there is inherent ambiguity when determining successful retrieval via recognition tests because "hits" are influenced by response bias as well as actual memory accuracy. This ambiguity in retrieval success is consistent with the substantial false alarm rates for new items exhibited in the two studies. Typically, recall tests are used for retrieval practice because they increase the power of the testing effect, while also providing a more direct measure of retrieval success (Rowland, 2014). A second issue with the use of recognition tests during retrieval practice is that each item is fully presented before a retrieval attempt can occur. This means that all items are re-experienced during phase 2, even if they are not successfully retrieved. Therefore, these retrieval practice conditions actually combined retrieval and restudy, making it difficult to attribute effects of DA to the encoding consequences of retrieval or to the encoding consequences of restudy (see Buchin & Mulligan, 2017, for additional detail on all of these points).

Because of the conflicting findings and issues discussed above, Mulligan and Picklesimer (2016) decided to assess the effects of DA on the mnemonic consequences of retrieval and restudy using the standard testing effect paradigm. The authors, in two experiments, had participants study word pairs (phase 1), practice retrieving the pairs on a cued-recall test or simply restudy the word pairs (phase 2), and then take an immediate or 1-day delayed final cued recall test (phase 3). Importantly, attention was manipulated during phase 2, so that half of the participants restudied or retrieved the word pairs under FA or DA.

Attention was divided using a concurrent task that required participants to classify aurally-presented digits as either even or odd. Performance on the final recall test revealed both a testing effect, such that retrieval practice produced better memory than restudy, and an effect of attention, such that FA during phase 2 led to better performance compared to DA during phase 2. Critically, the two factors interacted, such that DA during restudy was more detrimental to final performance compared to DA during retrieval. In turn, this interaction caused the testing effect to increase in size when phase 2 was conducted under DA. In other words, the relative benefits of testing increased under distraction.

Mulligan and Picklesimer (2016) interpreted this finding as support for the idea that both retrieval itself and the encoding effects of retrieval are resilient to distraction, possibly due to the obligatory nature of retrieval, compared to the sort of encoding that occurs in the restudy condition (Craik et al., 1996; Naveh-Benjamin et al., 2000). Their findings are also at odds with the predictions of the elaborative and effortful accounts of the testing effect, which propose that because the testing advantage stems from greater elaboration or effort, DA should negatively impact the benefits of retrieval practice.

Although the findings from Mulligan and Picklesimer (2016) are essential in determining how the encoding effects of retrieval operate, some important questions remain unanswered. First, does the relationship between the materials of the memory task and those of the secondary task affect the observed resiliency of the encoding effects of retrieval? Second, are the encoding effects of retrieval driven by obligatory retrieval processes or more controlled, strategic task prioritization? Recently, Buchin and Mulligan (2017) conducted three experiments designed to answer these questions.

To answer the first question, it is first helpful to understand the difference between material-general interference and material-specific interference (Fernandes & Moscovitch, 2000; see Mulligan, 2008, for review). The former refers to competition for general processing resources, produced when the materials of the dual-tasks are of different types (e.g., words for the memory task and digits for the secondary task). The latter refers to competition within the same representational system, produced when the materials of dual-tasks are drawn from the same category (e.g., both tasks use words). Prior research has shown that material specific-interference can produce more detrimental effects on memory retrieval compared to material-general interference (e.g. Fernandes & Moscovitch, 2000, 2002, 2003).

The experiments conducted by Mulligan and Picklesimer (2016) used different types of materials in the memory and secondary tasks: words in the former and digits in the latter. Therefore, their results demonstrated that material-general interference had a substantial effect on restudy but very little effect on the encoding consequences of retrieval. To answer the question of how the encoding effects of retrieval respond to material-specific interference, Buchin and Mulligan (2017) conducted two experiments that utilized similar materials for both the memory and distractor task. Participants first studied word pairs (phase 1) under FA and then restudied some of the word pairs (under FA or DA) or took a cued-recall test (under FA or DA; phase 2). Phase 3 consisted of a final, cued-recall test of all the word pairs. In the first experiment, the distractor task consisted of aurally-presented words that required immediate classification as either man- or nature-made. The second experiment also required the same classification, but the task was slowed and feedback was given after

each trial. Critically, the secondary task in both experiments used words, instead of digits, promoting material-specific interference (e.g. Fernandes & Moscovitch, 2000, 2002, 2003).

In general, the results replicated what was found by Mulligan and Picklesimer (2016) by revealing an interaction between attention and phase-2 learning condition (retrieval vs. restudy). As before, the testing effect increased in size under DA, compared to FA, demonstrating that the resilience of the encoding consequences of retrieval generalize over secondary-task materials (Buchin & Mulligan, 2017). Additionally, this finding appears inconsistent with the notion that the encoding effects of retrieval produce their benefits through enhanced elaboration or effort, which imply that DA should reduce this advantage.

The second question discussed above relates to secondary task performance and task prioritization. As noted earlier, secondary tasks typically impair encoding more than retrieval, but the costs to the secondary task are greater when paired with retrieval than encoding. This is part of the reason that Craik et al. (1996; Naveh-Benjamin et al., 2000) characterized retrieval as obligatory, arguing that it typically takes precedence over other ongoing activities, but not automatic, given that retrieval exacts large costs to ongoing secondary tasks. Additionally, manipulating task emphasis between the memory and secondary task produces large differential effects during encoding, but not retrieval, implying that encoding is under greater control than is retrieval (e.g. Craik et al., 1996). Overall, this suggests that the relatively obligatory nature of retrieval represents a fundamental difference between retrieval and encoding (Craik et al., 1996; Naveh-Benjamin et al., 2000; Naveh-Benjamin, Kilb, & Fisher, 2006), and one that might also be important to the encoding effects of retrieval.

However, the differences in secondary task performance between the encoding and retrieval conditions could simply reflect strategic task prioritization by the participants. Specifically, participants may be treating the retrieval task as more important (e.g., in the context of the digit-classification task) than the restudy task. If so, the mnemonic benefits of retrieval may arise from controlled, strategic processing, rather than an inherent obligatory retrieval process. This strategic variation in attentional allocation would be seen in poorer secondary task performance in the retrieval than restudy condition.

To assess this issue, I first return to the results of Mulligan and Picklesimer (2016), who replicated the usual phase 2 findings discussed above. The authors found that phase 2 retrieval success was minimally affected by the secondary task, producing only small and non-significant differences between the FA and DA conditions, but retrieval produced greater secondary task costs than did restudy. Specifically, secondary task performance was much more accurate in the restudy than retrieval conditions (i.e. 85% vs. 55% in experiment 1 and 87% vs. 64% in experiment 2). Although this phase 2 pattern was expected based on prior research, ambiguity remains regarding whether it is due to a necessary consequence of the obligatory nature of retrieval or due to differences in prioritization between encoding and retrieval in the face of distraction. To fairly assess the predictions of the elaborative and effortful accounts of the testing effect, strategic task prioritization must be assessed as a possible confounding factor.

Buchin and Mulligan (2017) attempted to resolve this ambiguity by designing two experiments (experiments 2 and 3) to produce high secondary task performance in both the phase 2 restudy and retrieval conditions. By modifying the speed of the secondary task and adding enhanced feedback, the authors hoped to minimize the usual differences in secondary

task performance between the restudy and retrieval conditions. Additionally, although the basic design of experiment 3 was the same as experiments 1 and 2, the secondary task used digits, instead of words. This manipulation was conducted in order to match the material-general interference caused by the materials of the secondary task in Mulligan and Picklesimer (2016). In general, both experiments found a similar pattern in final memory performance in that the word-pairs retrieved under DA during phase 2 were much less affected than word-pairs restudied under DA during phase 2, resulting in a larger testing effect in the DA compared to FA condition.

Returning to the secondary task, both experiments successfully increased secondary task performance and greatly reduced the difference in performance between restudy and retrieval. Specifically, the mean accuracies on the secondary task in the restudy and retrieval conditions were 87.8% and 84.4% (experiment 2) and 97% and 93% (experiment 3), respectively. These differences in secondary task performance are much smaller than those obtained in Mulligan and Picklesimer (2016) and suggest that the obligatory nature of retrieval, rather than strategic attentional allocation, drives the mnemonic benefit of retrieval (Buchin & Mulligan, 2017).

The results from the above experiments demonstrate the generality of the observed interaction between the phase 2 attention and learning condition. The pattern persists over shorter (a few minutes) and longer (24 hr) retention intervals (Mulligan & Picklesimer, 2016). It is found with related as well as unrelated word pairs, and whether retrieval practice uses feedback or not (Mulligan & Picklesimer, 2016). It occurs under material-general or material-specific interference (Buchin & Mulligan, 2017). Finally, the pattern is observed over higher or lower levels of secondary task performance, and whether the difference in

secondary task performance between the retrieval and restudy conditions is large or minimal (Buchin & Mulligan, 2017).

Overall, the results support the idea that the encoding effects of retrieval are largely resilient and obligatory in the face of distraction, like retrieval success itself. It should be noted, however, that this research used materials and methods typical of basic research in memory, methods which have limited ecological validity with respect to educational application. As stated earlier, much of the interest in the testing effect is motivated by its apparent applicability to real-world learning. In natural settings, learners are often challenged by varying amounts of distraction arising from external sources (e.g. the ambient distractions arising in the environment) and internal sources (e.g. mind wandering, and other off-topic internally-generated thoughts; Calderwood, Ackerman, & Conklin, 2014; Jacobesen & Forste, 2011; Smallwood, McSpadden, & Schooler, 2008; Szpunar, Khan, & Schacter, 2013). Consequently, it is important to examine the testing effect, and its relation to attention, with materials and methods more typical of educational settings.

To the author's knowledge, no study thus far has assessed the testing effect while using educational materials in conjunction with an attentional manipulation. However, researchers have examined the benefits of retrieval while using materials of the sort likely to appear in the classroom. For example, Carpenter et al. (2008) used Swahili-English word pairs to explore how retrieval practice benefits the learning of a foreign language. Because few participants already know Swahili, this type of material is common in memory studies that examine real-world language acquisition (e.g. Carpenter et al., 2008; Kang & Pashler, 2014; Karpicke & Roediger, 2008; Pyc & Rawson, 2009). The researchers found that retrieval practice resulted in better memory performance compared to restudy. Although

superficially similar to simple paired-associate tasks, learning novel foreign vocabulary is different (and more difficult; Caprneter et al., 2008) on several dimensions (e.g., mapping existing semantic representations onto a new symbol, learning new phonology, etc.). Consequently, it is important to know that the benefits of testing over restudy generalize to this type of real-world information.

Other educationally relevant materials and tests have been used in both laboratory and classroom experiments. Roediger and Karpicke (2006) asked participants to read short passages for a later memory test. Then, participants either reread the passage for additional study or recalled (and wrote down) all they could remember about the passage. On a final recall test, participants who practiced recalling the passage during phase 2 remembered more than participants who simply reread the passage. This testing effect was found in delayed final tests (2 days and 1 week) but not in an immediate test. Overall, these results show that recalling information from a passage leads to lower rates of forgetting on a final delayed test than rereading a passage (Roediger & Karpicke, 2006).

Classroom studies have revealed a similar memory benefit for retrieved compared to restudied materials (Butler & Roediger, 2007; McDaniel, Anderson et al., 2007). These studies found that short-answer practice tests were more effective than multiple-choice practice tests as well as simply restudying the material. Overall, testing effect studies that use more educationally-relevant materials and procedures replicate the usual finding that retrieval practice leads to better memory than restudy (Butler & Roediger, 2007; Carpenter et al., 2008; McDaniel, Anderson et al., 2007; Roediger & Karpicke, 2006). However, it remains to be seen if the mnemonic benefits of retrieval will remain resilient to DA while using materials and tests such as these.

To assess this question, two types of educational materials were used: Swahili-English word pairs (Experiments 1 and 2) and short prose passages (Experiment 3). Because these educationally-relevant materials are more complex than the simple word-pairs used previously (e.g. Buchin & Mulligan, 2017; Carpenter et al., 2008; Mulligan & Picklesimer, 2016), the detrimental effects of DA on memory may be more pronounced. Attention was divided using a digit classification task during the critical learning phase (i.e. phase 2) of the typical testing effect paradigm. Additionally, Experiment 2 was designed to assess the potential ambiguity discussed earlier regarding task prioritization between the restudy and retrieval conditions. Specifically, the continuous digit classification task was made easier in an attempt to equate secondary task performance between the two groups (similar to the task used by Buchin and Mulligan (2017), Experiment 3).

As with the prior experiments, both the elaborative and effortful accounts predict a decrease in the size of the testing effect under DA, compared to FA. Because elaborative and effortful processing are both susceptible to distraction (Craik et al., 1996; Mulligan, 2008), dividing attention during retrieval practice should reduce the mnemonic benefits of retrieval. However, if the encoding effects of retrieval remain resilient to distraction, like retrieval success itself, then dividing attention during retrieval practice should have little impact on later memory performance. Consequently, this would lead to an unchanged or larger testing effect in the DA, compared to the FA, condition. Although prior research suggests the latter pattern of results (i.e. Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016), the complexity of educational materials may increase the reliance of retrieval on attentional resources.

CHAPTER 2: EXPERIMENT 1

Experiment 1 used Swahili-English word pairs and cued recall for both retrieval practice and for the final memory test. Phase 1 consisted of presenting the full word pairs for initial study. In phase 2, the word pairs were presented again for additional restudy or for retrieval practice. During retrieval practice, only the first (cue) word of the pair was presented and the participants were instructed to recall the second (target) word. Regardless of the participant's answer, the full cue-target pair was shown afterwards as feedback. During phase 2, the word pairs were either presented once or three times, although the learning condition (i.e. restudy or retrieval) was the same for each of the repeated presentations. Prior research on attention and the testing effect used only one presentation during phase 2 (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016), however the increased difficulty of foreign language learning may require repeated restudying and retrieving. Therefore, repetition was varied such that half the pairs were presented once for restudy or retrieval during phase 2 (1-rep) and half were presented three times during phase 2 (3-rep).

Additionally, phase 2 took place either under FA or DA conditions. Attention was divided using a continuous digit classification task, previously used by Mulligan and Picklesimer (2016). Specifically, this task requires classifying aurally-presented digits as either even or odd. The third phase of the experiment took place two days later and consisted of a cued recall test in which the first word from each pair was presented as a cue for the recall of the second (target) word. A two-day delay was used to mimic studying and testing

in the real world and because a testing effect does not always emerge with an immediate final test (e.g. Roediger & Karpicke, 2006).

As stated above, if the encoding effects of retrieval are similar to typical encoding processes, one might expect a marked reduction in final recall in the DA retrieval condition, perhaps equal to the expected reduction in final recall for the DA restudy condition. In contrast, the elaborative and effortful accounts of the testing effect go even further, predicting that the effects of DA should be greater in the retrieval than restudy condition. This in turn predicts a smaller testing effect under DA compared to FA. Alternatively, the mnemonic benefits of retrieval may remain resilient to distraction, even when using a more complex memory task (i.e. foreign-language learning). If so, this would replicate the findings by Buchin and Mulligan (2017) and Mulligan and Picklesimer (2016) who reported that the testing effect became significantly larger under DA, compared to FA.

Methods

Participants. Forty participants from UNC at Chapel Hill were recruited in exchange for course credit. The sample size was chosen based on past results from this line of research as well as on prior testing-effect studies using educational materials. In particular, the effect size of the interaction between phase 2 condition and attention from Buchin and Mulligan (2017) averaged $d = .784$. For this effect size, 24 participants in the FA and DA groups are required to yield power of .95 (Faul, Erdfelder, Lang, & Buchner, 2007). The sample size in prior testing-effect research using educationally relevant materials has typically ranged from 20 to 48 participants in each critical condition (i.e. Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler, Karpicke, & Roediger, 2008; Carpenter, 2011; Carpenter et al., 2008; Pyc & Rawson, 2009, 2010, 2011). Therefore, 40 participants were chosen as a

compromise between the power analysis and the prior research. The study received research ethics committee (Instructional Review Board) approval

Materials and Design. The memory task used sixty-four Swahili-English word pairs (e.g. somo-friend, farasi-horse, gereza-jail) that were assembled from Carpenter et al. (2008) and Pyc and Rawson (2009). According to Wilson's (1988) norms, the English words were all nouns that ranged between three and seven letters, one and three syllables, occurred with a frequency of greater than 20 per million, and ranged in concreteness from 400 to 700. The Swahili equivalents were originally obtained from the Kamusi Project Web site (Yale University, 2005) as well as from Nelson and Dunlosky (1994).

Phase 2 learning condition (restudy vs. retrieval), attention (full vs. divided), and repetition (1-rep vs. 3-rep) were manipulated within-subjects. The word pairs were randomly divided into eight sets of eight and assigned to the eight possible condition combinations (FA-restudy 1-rep, FA-restudy 3-rep, FA-retrieval 1-rep, FA-retrieval 3-rep, DA-restudy 1-rep, DA-restudy 3-rep, DA-retrieval 1-rep, DA-retrieval 3-rep). Phase 2 was organized into four acquisition blocks, one each for the FA-restudy, FA-retrieval, DA-restudy, and DA-retrieval conditions. The order of the blocks was counterbalanced across participants. Within each block, 8 word pairs were presented once (1-rep) and 8 word pairs were presented three times (3-rep) in an intermixed order, totaling 32 trials per block. The items were ordered such that the average list position of the word pairs presented once (1-rep) was equal to the average list position of the word pairs presented three times (3-rep). Critically, each pair was presented in only one block, and so was experienced in only one of the learning-condition-and-attention combinations. Additionally, attention was divided using a secondary task

consisting of randomly presented digits (1-9) played over the headphones which participants categorized as even or odd.

Procedure. The experiment consisted of three phases. Phase one was a study phase in which participants were shown each of the 64 Swahili-English word pairs. Each study trial began with a 250 ms blank screen, followed by the full cue-target word pair for 6,000 ms. There was no distraction during this initial study phase and each pair was presented a single time. Before beginning the study phase, participants were told that they should try to learn the pairs for a later memory test in which the first word from each pair will be presented and the second word should be recalled.

The second phase consisted of presenting the word pairs for additional learning under four consecutive acquisition blocks that each began immediately after the prior block was completed. A block consisted of 16 word pairs that were each presented either once (1-rep) or three times (3-rep), totaling 32 trials per block. Participants were told that some of the word pairs would be repeated within a block and some pairs would be presented only once.

Specifically, in the FA-restudy block, participants were told they would see full cue-target pairs and were instructed to read the first word to themselves and the second word out loud (this vocal response was necessary to equate the overt response in the restudy and retrieval conditions). Each trial began by displaying a word pair on the computer screen for 6,000 ms, preceded by a blank screen for 250 ms. In the FA-retrieval block, participants were shown only the cue word and told to read it to themselves before recalling the target word out loud. For each trial, the cue word was presented for 4,500 ms and then the full cue-target word pair was shown for 1,500 ms, followed by a blank screen for 250 ms. The participant

was instructed to try to recall the answer in the first 4,500 ms, and then look at the feedback, saying the correct answer aloud if they either failed to answer or recalled an incorrect answer.

The DA-restudy and DA-retrieval blocks were the same as their FA counterparts, with the addition of the secondary task. Participants were told that the secondary task and restudying/retrieving the word pairs (i.e. the memory task) were equally important. During the DA blocks, participants heard a series of digits (1-9) over the headphones, at a rate of one digit every 1500 ms, for a total of 4 digits per memory trial (the first digit was presented as the word pair appeared on the computer screen). Participants were instructed to listen to the digits and use the keyboard to indicate if the digit was odd (i.e. press the "o" button) or even (i.e. press the "e" button). If the answer was correct, no feedback was given, but if the participant responded incorrectly or took too long to respond (i.e. no response after 1500 ms), they heard a buzzer noise.

After completing all four phase 2 acquisition blocks, participants left the lab and were instructed to return two days later. Upon returning, participants began the final cued-recall test for all of the Swahili-English word pairs (phase 3). Each cue word was presented for 6 s. followed by a blank screen for 250 ms. Participants were instructed to recall aloud the English word associated with each Swahili word presented within the 6 seconds.

Results

Phase 2 Cued Recall. The proportion of targets recalled during phase 2 (Table 1) was analyzed with a 2 x 2 ANOVA, using phase 2 attention (FA vs. DA) and phase 2 repetition (1-rep vs. 3-rep) as within-subjects factors. Phase 2 recall for the 3-rep condition represents the average recall over all three retrieval attempts. First, a main effect of attention was found, $F(1, 39) = 30.103$, $MS_e = .015$, $p < .001$, $\eta^2_p = .436$, showing greater recall in the

FA than DA conditions. Second, a main effect of repetition was also found, $F(1, 39) = 112.872$, $MS_e = .014$, $p < .001$, $\eta^2_p = .743$, indicating higher recall rates for pairs that were repeated three times during phase 2. Third, the interaction between phase 2 attention and repetition was also significant, $F(1, 39) = 10.967$, $MS_e = .011$, $p = .002$, $\eta^2_p = .219$.

Follow-up tests were conducted to further investigate the significant interaction. Dividing attention during phase 2 significantly reduced recall in the 3-rep condition, $t(39) = 6.544$, $SE = .025$, $p < .001$, $d = 0.937$, and was trending towards significance in the 1-rep condition, $t(39) = 2.01$, $SE = .026$, $p = .051$, $d = 0.362$. As one would expect, repetition enhanced recall under FA, $t(39) = -9.455$, $SE = .026$, $p < .001$, $d = 1.467$, and under DA, $t(39) = -6.17$, $SE = .023$, $p < .001$, $d = 0.94$.

Phase 3 Cued Recall. The critical final recall results were measured as the proportion of target words recalled[1]. The final recall scores were analyzed with a 2 x 2 x 2 ANOVA, using phase 2 attention (FA vs. DA), phase 2 condition (restudy vs. retrieval), and phase 2 repetition (1-rep vs. 3-rep) all as within-subjects factors (Figure 1). All three main effects were significant, indicating: (1) a detrimental effect of phase 2 DA, $F(1, 39) = 68.09$, $MS_e = .02$, $p < .001$, $\eta^2_p = .636$; (2) a benefit of phase 2 retrieval practice (i.e. a testing effect), $F(1, 39) = 35.416$, $MS_e = .011$, $p < .001$, $\eta^2_p = .476$; and (3) a benefit of phase 2 repetition, $F(1, 39) = 43.804$, $MS_e = .019$, $p < .001$, $\eta^2_p = .529$. Additionally, two of the three two-way interactions were significant, specifically: (1) the interaction between phase 2 attention and repetition, $F(1, 39) = 8.626$, $MS_e = .015$, $p = .006$, $\eta^2_p = .181$, indicating that the negative effect of DA was greater in the 3-rep than 1-rep condition; (2) the interaction between phase

[1] In testing-effect research, another measure of final recall is also often used – the conditionalized recall score. This measure was calculated as the proportion of target words recalled out of only those words successfully retrieved at least once during phase 2. For completeness, analyses using this measure are presented in Appendix A. For all 3 experiments, these additional analyses did not change any of the primary conclusions of the paper with respect to the effects of attention on the encoding consequences of restudy and retrieval.

2 condition and repetition, $F(1, 39) = 6.214$, $MS_e = .026$, $p = .017$, $\eta^2_p = .137$, indicating that the testing effect was larger in the 3-rep than 1-rep condition. The interaction between phase 2 attention and condition was non-significant, $F(1, 39) = 0.024$, $p = .878$, as was the three-way interaction, $F(1, 39) = 0.317$, $p = .577$.

The significant two-way interactions both involved the repetition condition and so are not of primary concern. Regardless, further analyses were conducted and are briefly discussed below. First, the significant interaction between phase 2 attention and repetition indicates that the disruptive effect of DA was greater in the 3-rep than 1-rep condition. To fully evaluate this interaction, a separate 2 (attention) x 2 (phase 2 condition) ANOVA was conducted for each level of phase 2 repetition (1-rep vs. 3-rep). For the 1-rep condition, the effect of attention was significant, $F(1, 39) = 28.289$, $MS_e = .012$, $p < .001$, $\eta^2_p = .42$, the testing effect was non-significant, $F(1, 39) = 2.353$, $p = .133$, and the interaction was non-significant, $F(1, 39) = 0.067$, $p = .797$. For the 3-rep condition, the effect of attention was significant, $F(1, 39) = 50.961$, $MS_e = .023$, $p < .001$, $\eta^2_p = .566$, the testing effect was significant, $F(1, 39) = 21.434$, $MS_e = .025$, $p < .001$, $\eta^2_p = .355$, and the interaction was non-significant, $F(1, 39) = 0.198$, $p = .659$. Taken together, the results indicate that the disruptive effect of DA was larger in the 3-rep, than 1-rep, condition, although the effect was significant in both.

The attention-by-repetition interaction can also be interpreted as showing that the effects of repetition increased under FA compared to DA. Similar to the prior analyses, a 2 (phase 2 condition) x 2 (repetition) ANOVA was conducted at each level of phase 2 attention (FA vs. DA). For the FA condition, the testing effect was significant, $F(1, 39) = 12.538$, $MS_e = .017$, $p = .001$, $\eta^2_p = .243$, the effect of repetition was significant, $F(1, 39) = 38.089$, $MS_e =$

.021, $p < .001$, $\eta^2_p = .494$, and the interaction was significant, $F(1, 39) = 4.285$, $MS_e = .025$, $p = .045$, $\eta^2_p = .099$. For the DA condition, the testing effect was significant, $F(1, 39) = 15.029$, $MS_e = .013$, $p < .001$, $\eta^2_p = .278$, the effect of repetition was significant, $F(1, 39) = 12.381$, $MS_e = .013$, $p < .001$, $\eta^2_p = .241$, and the interaction was significant, $F(1, 39) = 4.267$, $MS_e = .013$, $p = .046$, $\eta^2_p = .099$. Therefore, although the benefit of repetition was larger in the FA than DA condition, the effect of repetition was significant in both attention conditions.

Finally, the significant two-way interaction between phase 2 condition and repetition indicates that the testing effect was larger in the 3-rep than 1-rep condition. The prior analyses found that the testing effect was significant only in the 3-rep but not 1-rep condition. This can be clearly seen in Figure 2, which plots the testing effect (the difference in recall between the retrieval and restudy conditions) as a function of attention and repetition. The condition-by-repetition interaction can also be interpreted as an increased benefit of phase 2 repetition for words retrieved compared to words restudied. To investigate this interpretation, a 2 (attention) x 2 (repetition) ANOVA was conducted at each level of phase 2 condition (restudy vs. retrieval). For the restudy condition, the effect of attention was significant, $F(1, 39) = 29.105$, $MS_e = .023$, $p < .001$, $\eta^2_p = .427$, the effect of repetition was significant, $F(1, 39) = 6.273$, $MS_e = .021$, $p = .017$, $\eta^2_p = .139$, and the interaction was non-significant, $F(1, 39) = 3.162$, $p = .083$. For the retrieval condition, the effect of attention was significant, $F(1, 39) = 45.901$, $MS_e = .016$, $p < .001$, $\eta^2_p = .541$, the effect of repetition was significant, $F(1, 39) = 36.959$, $MS_e = .023$, $p < .001$, $\eta^2_p = .487$, and the interaction was significant, $F(1, 39) = 6.476$, $MS_e = .014$, $p = .015$, $\eta^2_p = .142$. Taken together, the results indicate that although the benefit of repetition during phase 2 was larger in the retrieval than restudy condition, the effect of repetition was significant in both conditions.

Phase 2 Digit Classification Task. The proportion of correctly classified digits during phase 2 (Table 2) was analyzed using a 2 (phase 2 condition) x 2 (repetition) ANOVA. Only the main effect of condition was significant (other F's < 1), indicating greater accuracy when paired with restudy than retrieval, $F(1, 39) = 186.549$, $MS_e = .008$, $p < .001$, $\eta^2_p = .827$. Although it may seem like digit task accuracy was near chance in the retrieval condition, this is due to omissions rather than incorrect responses. To investigate this, the same 2 x 2 ANOVA was used to analyze only those trials with actual responses (i.e. excluding omissions), which increased the proportion of correct responses in both conditions (Table 2). Specifically, the proportion correct in the restudy condition increased by about 10% and the retrieval condition increased by about 20%. Despite the larger increase for the retrieval condition, the main effect of phase 2 condition remained significant, $F(1, 39) = 65.12$, $MS_e = .003$, $p < .001$, $\eta^2_p = .625$, again indicating higher performance on the digit task when paired with restudy, than with retrieval. As with the unrestricted analysis above, only the main effect of condition was significant (other $Fs < 1.5$).

Discussion

The critical phase 3 recall results will be discussed first, before turning to the phase 2 recall and secondary task results. Overall, retrieval practice led to greater final recall than restudy (i.e. a testing effect), DA during phase 2 disrupted final recall compared to FA, and word pairs that were repeated during phase 2 had higher rates of final recall than those that were not. However, the main focus of Experiment 1 was to assess how DA affects the encoding consequences of restudy and retrieval. The interaction between phase 2 attention and condition was non-significant, indicating similar levels of disruption due to DA between phase restudy and retrieval.

Because the attention manipulation had the same effect on the retrieval and restudy condition, the size of the testing effect was similar under DA and FA, which does not support the predictions of the effortful and elaborative accounts of the testing effect. Specifically, both effortful and elaborative processing are attention demanding (i.e. Craik & Broadbent, 1983; Craik et al., 1996; Hasher & Zacks, 1979; McDowd & Craik, 1988; Mulligan, 2008), which indicates that if the encoding effects of retrieval did depend on greater effort or elaboration, then the disruption caused by DA during phase 2 would be greater for retrieval than restudy.

The lack of support for the predictions of the effortful and elaborative accounts of the testing effect is in line with prior research (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016), but the specific pattern of results does differ. Specifically, prior research found that the disruptive effect of DA was significant in the restudy condition but non-significant in the retrieval condition. Alternatively stated, the testing effect was larger in the DA than FA condition (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). However, the results of Experiment 1 revealed testing effects of similar size regardless of attention condition. This difference between experiments may reflect the importance of the memory task materials. The complexity and difficulty of these educational materials (e.g. Carpenter et al., 2008) may have increased the reliance of retrieval on attentional resources (I return to this issue in the General Discussion).

The two significant two-way interactions also warrant brief discussion. First, phase 2 DA disrupted final recall in both the 1-rep and 3-rep conditions, although the disruption was larger for pairs repeated during phase 2. Second, the relative benefit of retrieval practice compared to restudy (i.e. the testing effect) was significant in the 3-rep, but not 1-rep,

condition. Third, although the benefit of repetition was larger in the FA, than DA, condition as well as in the retrieval, than restudy, condition, the effect was significant in all possible condition combinations. This pattern likely reflects the difficulty in novel foreign language learning, especially if given only one restudy or retrieval practice opportunity. It is likely that these interactions are at least partly due to the relatively low recall in the 1 rep condition.

Although the phase 2 results are not the primary concern of the current study, two aspects of the results deserve mention. First, phase 2 recall performance was significantly reduced under DA compared to FA. Although DA during retrieval is usually less disruptive than DA during encoding, it can produce modest, sometimes significant, reductions in retrieval success (e.g. Craik et al., 1996). Additionally, negative effects of DA on memory generally increase as task difficulty increases (e.g. McDowd & Craik, 1988) and the memory task used in the current experiment (i.e. recall of Swahili-English word pairs) has been characterized as difficult (e.g. Carpenter et al., 2008). This may be why significant reductions in phase 2 recall were found in the present experiment. Second, performance on the secondary task was worse in the retrieval than restudy condition, consistent with the typical finding that secondary-task costs are greater for retrieval than encoding (or re-encoding) (e.g., Craik et al., 1996).

CHAPTER 3: EXPERIMENT 2

Prior research on attention and memory has demonstrated differential effects on secondary task performance when attention is divided during encoding and retrieval (e.g. Craik et al., 1996). Specifically, secondary task performance is more greatly affected by retrieval compared to restudy. Consequently, it was expected that Experiment 1 would produce a similar pattern in secondary task performance. Not only was this pattern expected, but it also appears to reflect participant's natural inclinations when confronted with a dual-task situation. Given that this set of experiments is designed to emphasize real-world learning and ecological validity, it is important to let participants behave in an unconstrained manner when coordinating multiple cognitive demands.

Although important for ecological validity, differences in secondary task performance introduce a potential ambiguity in interpreting the results. That is, this pattern may reflect the obligatory nature of retrieval or it might reflect task prioritization differences between encoding and retrieval (an ambiguity discussed in the introduction; see Buchin and Mulligan, 2017, for more detail). Therefore, because Experiment 1 revealed the usual pattern of secondary task performance (e.g., Craik et al., 1996), Experiment 2 was designed to better equate performance on the secondary task between restudy and retrieval without sacrificing ecological validity. One possible method requires participants to monitor their performance via enhanced feedback (ex. displaying cumulative secondary task accuracy as a running percentage on the screen) in order to maintain a specific level of accuracy (ex. above 75%). Although this method would help decrease differences in secondary task performance

between restudy and retrieval, the results would no longer reflect participant's natural inclinations when coordinating multiple cognitive demands. Therefore, in a continued effort to emphasize real-world learning, a different method was used to better equate secondary task performance.

To achieve this, the secondary task was changed from a continuous classification task to a somewhat easier, cumulative classification task. This new task was based on the secondary task used by Buchin and Mulligan (2017, Experiment 3), which minimized differences in secondary task performance between restudy and retrieval. Specifically, instead of requiring participants to immediately classify digits, the task now required them to keep track of the number of odd digits presented during a single trial. This task also features enhanced feedback as described below. All other aspects of the experiment were the same as Experiment 1.

Methods

Participants. Forty participants from UNC at Chapel Hill were recruited in exchange for course credit.

Materials, Design, and Procedure. The methods were the same as Experiment 1 except for the following modifications to the secondary task. During the phase 2 DA blocks, participants now had to keep track of the number of odd digits they heard during the trial. Specifically, during these trials, participants heard 4 digits (drawn from the set 1-9) over the headphones during the 6000 ms phase 2 trial, at a rate of one digit every 1500 ms, with the first digit being played as the word-pair was first presented. Participants were instructed to keep track of the number of odd digits presented during the trial. At the end of the trial, the screen displayed "how many of the digits were odd?" and prompted participants to enter their

response (0 – 4) using the keyboard. If the response was correct, the screen displayed "Correct!" and the participant would press the space bar to advance to the next trial. If the response was incorrect, the screen displayed "Incorrect!" and a buzzer sound played over the headphones before the participant pressed the space bar to advance to the next trial. This change was designed to better equate performance on the secondary task between the restudy and retrieval conditions while still allowing participants to behave in an unconstrained manner. All other aspects of Experiment 2 were identical to Experiment 1.

Results

Phase 2 Cued Recall. As in Experiment 1, the proportion of correctly recalled targets during phase 2 (Table 1) was analyzed using a 2 x 2 ANOVA, with phase 2 attention (full vs. divided) and phase 2 repetition (1-rep vs. 3-rep) as within-subjects factors. Replicating Experiment 1, all effects were significant, specifically: (1) a main effect of attention, indicating greater recall under FA, than DA, $F(1, 39) = 11.461$, $MS_e = .015$, $p = .002$, $\eta^2_p = .227$; (2) a main effect of repetition, revealing higher recall for repeated pairs, than pairs presented only once during phase 2, $F(1, 39) = 17.372$, $MS_e = .021$, $p < .001$, $\eta^2_p = .548$; and (3) an interaction between phase 2 attention and repetition, $F(1, 39) = 27.137$, $MS_e = .011$, $p < .001$, $\eta^2_p = .41$. Follow-up tests indicated that the effect of attention was significant for the repeated pairs, $t(39) = 6.154$, $SE = .025$, $p < .001$, $d = 0.92$, but not for pairs presented only once, $t(39) = -0.713$, $p = .48$. Additionally, the benefit of repetition was found under FA, $t(39) = -9.025$, $SE = .027$, $p < .001$, $d = 1.578$, as well as under DA, $t(39) = -2.512$, $SE = .029$, $p = .016$, $d = 0.521$.

Phase 3 Cued Recall. As in Experiment 1, the final recall scores were analyzed with a 2 x 2 x 2 ANOVA, using phase 2 attention (FA vs. DA), phase 2 condition (restudy vs.

retrieval), and phase 2 repetition (1-rep vs. 3-rep) all as within-subjects factors (Figure 3). All main effects were significant indicating that: (1) DA during phase 2 disrupted later recall compared to FA, $F(1, 39) = 20.075$, $MS_e = .022$, $p < .001$, $\eta^2_p = .34$; (2) retrieval practice during phase 2 enhanced final recall compared to restudying (i.e. a testing effect), $F(1, 39) = 18.314$, $MS_e = .028$, $p < .001$, $\eta^2_p = .32$; and (3) repetition during phase 2 benefited later memory, $F(1, 39) = 61.478$, $MS_e = .02$, $p < .001$, $\eta^2_p = .612$. The two-way interaction between phase 2 attention and repetition was also significant, $F(1, 39) = 13.886$, $MS_e = .014$, $p = .001$, $\eta^2_p = .263$, indicating that the disruptive effect of phase 2 DA was greater in the 3-rep than 1-rep condition. The two-way interaction between phase 2 condition and repetition approached significance, $F(1, 39) = 3.647$, $MS_e = .026$, $p = .064$, $\eta^2_p = .086$, suggesting that the testing effect was larger in the 3-rep than 1-rep condition. All other interactions were non-significant (ps > .05).

To fully assess the attention-by-repetition interaction, a 2 (attention) x 2 (phase 2 condition) ANOVA was conducted at each level of phase 2 repetition (1-rep vs. 3-rep). For the 1-rep condition, the effect of attention was non-significant, $F(1, 39) = 1.786$, $p = .189$, the testing effect was significant, $F(1, 39) = 4.457$, $MS_e = .018$, $p = .041$, $\eta^2_p = .103$, and the interaction was non-significant, $F(1, 39) = 0.225$, $p = .638$. For the 3-rep condition, the effect of attention was significant, $F(1, 39) = 30.277$, $MS_e = .02$, $p < .001$, $\eta^2_p = .437$, the testing effect was significant, $F(1, 39) = 14.77$, $MS_e = .035$, $p < .001$, $\eta^2_p = .275$, and the interaction was non-significant, $F(1, 39) = 0.239$, $p = .628$. The results indicate that phase 2 DA significantly disrupted later recall in the 3-rep but not 1-rep condition.

The phase 2 attention and repetition interaction can also be interpreted as indicating a larger benefit of repetition under FA than DA. To assess this, a 2 (phase 2 condition) x 2

(repetition) ANOVA was conducted at each level of phase 2 attention (FA vs. DA). For the FA condition, the testing effect was significant, $F(1, 39) = 10.16$, $MS_e = .031$, $p = .003$, $\eta^2_p = .207$, the effect of repetition was significant, $F(1, 39) = 76.528$, $MS_e = .016$, $p < .001$, $\eta^2_p = .662$, and the interaction was non-significant, $F(1, 39) = 2.049$, $p = .16$. For the DA condition, the testing effect was significant, $F(1, 39) = 13.13$, $MS_e = .015$, $p = .001$, $\eta^2_p = .252$, the effect of repetition was significant, $F(1, 39) = 12.935$, $MS_e = .018$, $p = .001$, $\eta^2_p = .249$, and the interaction was non-significant, $F(1, 39) = 3.308$, $p = .077$. Taken together, the results show that although the benefit of phase 2 repetition was significant across all conditions, it was larger under FA than DA.

Phase 2 Digit Classification Task. The proportion of correct answers on the digit classification task (Table 2) was analyzed using a 2 (phase 2 condition) x 2 (repetition) ANOVA. Similar to the results of Experiment 1, only the main effect of condition was significant (other $Fs < 1$), revealing greater accuracy in the secondary task when paired with restudy than retrieval, $F(1, 39) = 57.34$, $MS_e = .013$, $p < .001$, $\eta^2_p = .595$. Because of the methodological change in the secondary task, all trials now required responses, which renders a separate analysis excluding omissions (as in Experiment 1) unnecessary.

Discussion

The phase 3 recall results in Experiment 2 were very similar to Experiment 1. First, a testing effect was found, indicating higher final recall due to retrieval practice than restudy. Second, phase 2 DA significantly reduced final recall compared to phase 2 FA. Third, pairs that were repeatedly presented during phase 2 had greater final recall than pairs presented just once. Critically, the interaction between phase 2 attention and condition was non-significant, replicating Experiment 1. Thus, the encoding consequences of restudy and

retrieval were similarly disrupted by DA. Alternatively stated, the size of the testing effect was similar between the phase 2 FA and DA conditions. This can be seen in Figure 4, which plots the size of the testing effect as a function of attention and repetition.

Similar to prior research (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016) and Experiment 1, the testing effect did not decrease in size under DA. Therefore, the encoding effects of retrieval, as compared to restudy, do not seem to rely on greater effortful or elaborative processing, which are both highly susceptible to DA (e.g., Craik & Broadbent, 1983; Craik et al., 1996; Hasher & Zacks, 1979; McDowd & Craik, 1988; Mulligan, 2008). However, the testing effect also did not increase in size under DA, compared to FA, as it did in prior research (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). Experiment's 1 and 2 in the current study both used novel foreign language word pairs and both found a significant effect of DA on the retrieval condition, whereas the effect was non-significant in the prior research (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). One reasonable speculation consistent with these results concerns the complexity of the memory task materials. Specifically, the complexity of the current materials (e.g., Carpenter et al., 2008) may have increased attentional demands during retrieval practice to a greater degree than the materials used in prior studies (i.e. Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016; I return to this issue in the General Discussion).

Similar to Experiment 1, the two-way interaction between phase 2 attention and repetition was significant and subsequently assessed through additional 2 x 2 ANOVA's. First, phase 2 DA significantly disrupted final recall in the 3-rep, but not 1-rep, condition. Second, the benefit of phase 2 repetition was significant in both attention conditions, although it was much larger in the FA, than DA, condition. As in Experiment 1, the

interaction is likely due to the relatively low level of recall in the 1-rep condition. Consequently, because final recall in the 1-rep condition was near floor, the disruptive effects of phase 2 DA may not have had room to emerge.

Lastly, the phase 2 recall and digit task performance results are discussed. The pattern of the phase 2 recall results was similar to Experiment 1, except for a slight difference in the 1 rep condition. Specifically, in this condition, DA had a marginally significant effect in Experiment 1 (p = .051) whereas the effect was non-significant in Experiment 2 (p = .48). The current DA task might be somewhat easier than in Experiment 1, which may explain the slight discrepancy in results. But more generally, the finding that DA generally disrupted phase 2 recall may also reflect the enhanced difficulty of the current memory task (e.g. Carpenter et al., 2008; McDowd & Craik, 1988).

More relevant to Experiment 2 is that performance on the digit classification task continued to suffer to a greater degree when paired with retrieval than with restudy. Although the obligatory and attention-demanding nature of retrieval is well documented (e.g., Anderson et al., 1998; Craik et al., 1996; Naveh-Benjamin et al., 2000) and in line with the results of Experiment 1, it is still interesting that the modifications to the secondary task did not fully equate digit task performance between restudy and retrieval. However, the modifications did result in two important differences between the experiments and between the conditions. First, digit task performance between restudy and retrieval was better equated in Experiment 2 than 1 (i.e. the difference in digit task accuracy between restudy and retrieval was smaller in Experiment 2 than 1). Second, digit task accuracy in the retrieval condition benefited from the secondary task modification to a greater degree than did the

restudy condition (i.e. the difference in digit task accuracy between Experiments 1 and 2 was larger in the retrieval than restudy condition)[2].

Overall, the attention-demanding nature of retrieval continued to impact secondary task performance to a greater degree than restudy, although the modifications in Experiment 2 did reduce that difference. Importantly, secondary task performance was better equated without coming at a large cost to ecological validity (i.e. participants' behavior was much less constrained than it would be if a specific level of accuracy was required and constantly monitored).

[2] Digit task accuracy in the restudy and retrieval condition was compared between experiments using a 2 x 2 x 2 ANOVA, with phase 2 condition and repetition as within-subjects factors and experiment number (1 vs. 2) as a between-subjects factors. Only three effects were significant (other F's < 1.1), including the typical main effect of condition, $F(1, 78) = 207.677$, $MS_e = .069$, $p < .001$, $\eta^2_p = .727$, as well as the main effect of experiment, $F(1, 78) = 21.783$, $MS_e = .044$, $p < .001$, $\eta^2_p = .218$, indicating higher digit task accuracy in Experiment 2 than 1. Critically, the interaction between condition and experiment was also significant, $F(1, 78) = 6.406$, $MS_e = .0.011$, $p = .013$, $\eta^2_p = .076$, which indicates that the effect of condition was smaller in Experiment 2, than 1. Follow-up tests assessing the size of the restudy advantage revealed a marginally significant difference between experiments in the 1-rep condition, $t(78) = 1.876$, $SE = .035$, $p = .064$, $d = .419$, and a significant difference in the 3-rep condition, $t(66.311) = 2.061$, $SE = .025$, $p = .043$, $d = .461$. Thus the modification to the secondary task in Experiment 2 did better equate performance between restudy and retrieval.

CHAPTER 4: EXPERIMENT 3

The main purpose of Experiment 3 was to generalize the results of the prior experiments to other materials that are educationally relevant and even more complex. Experiment 3 used prose passages instead of foreign-language word pairs as the memory task material. In phase 1, participants read a passage of educational text (under full attention). In phase 2, participants either studied seven key concepts extracted from the passage (restudy condition) or were presented with seven similar fill-in-the-blank short answer questions (retrieval condition). As before, phase 2 was conducted either under FA or DA. Attention was divided using the same continuous digit task used in Experiment 1, which was chosen over the cumulative digit task used in Experiment 2 for two reasons. First, performance on the cumulative task was not fully equated between the restudy and retrieval conditions. Second, the cumulative task would require participants to keep track of more digits per trial than in Experiment 2 (4 vs. 9) at the same time as they read sentences, instead of word pairs. This would increase the difficulty of the cumulative task and likely eliminate (or even reverse) its advantage relative to the continuous digit task. Participants repeated phases 1 and 2 with four different educational passages, such that each of the possible combinations of attention and phase 2 conditions was experienced. Participants then left the lab and returned two days later for the final tests for each of the passages.

Methods

Participants. Forty participants from UNC at Chapel Hill were recruited in exchange for course credit.

Materials and Design. Four prose passages were adapted from Agarwal et al., (2008). The passages originally came from an educational textbook (Cooper et al., 1996). Each passage was approximately 1000 words in length (M = 1008), covered a single topic ('Earthquakes', 'Fossils', 'Voyager', and 'Wolves'), and the average Flesch Reading Ease score for the four passages was 71.4 (Flesch, 1948). Each passage also had seven corresponding fill-in-the-blank, short answer questions based on facts and ideas contained in each passage[3]. The fill-in-the-blank questions were adapted from the short answer, free-response questions used by Agarwal et al. (2008). For example, the following excerpt is from the 'Earthquakes' passage:

> '… Sand sometimes bubbles up during earthquakes, gushing water and soil like miniature mud volcanoes. These "sand boils" are particularly dangerous to buildings. In places where water is close to the surface, sandy layers turn into quicksand and structures crumble. …'

The corresponding fill-in-the-blank question asked, 'Miniature mud volcanoes that bubble up during earthquakes are called _____'. The answer was, "sand boils." Each fill-in-the-blank question had a corresponding restudy sentence with the correct answer in place of the blank, for example, 'Miniature mud volcanoes that bubble up during earthquakes are called sand boils'. The test questions were identical on the initial and final tests, similar to the format used in prior studies assessing the testing effect with educational materials (e.g. Agarwal et al., 2008; Kang et al., 2007; McDaniel, Anderson et al., 2007; Roediger, &

[3] The fill-in-the-blank, short answer questions were pilot tested to determine if participants could correctly answer the question without having read the passage (that is, on the basis of their prior general knowledge). Only 3 questions were answered correctly by more than 10% of the pilot subjects. Those three questions were replaced and the set was re-assessed with new subjects. Of the final set of 28 questions, 26 questions were never answered correctly and 2 questions were correctly answered by only one participant each. Thus the rate of correctly answering the final set of questions without having first read the passages was 0.84%, an acceptably low baseline rate.

Karpicke, 2006; see Wooldridge, Bugg, McDaniel, & Liu, 2014 for a cautionary note). Questions appeared on each test in the order in which the facts/concepts occurred in the passage.

As in Experiment 1 and 2, phase 2 learning condition (restudy vs. retrieval) and attention during phase 2 (full vs. divided) was manipulated within-subjects and counterbalanced for order. Unlike in the prior experiments, there was no phase 2 repetition condition because each key concept was either restudied or tested once during phase 2. Each set of seven sentences corresponding to one passage was assigned (and counterbalanced) to each of the four possible condition combinations (FA-restudy, FA-retrieval, DA-restudy, and DA-retrieval). As in Experiment 1, attention was divided in the two DA conditions by requiring participants to continually classify digits (1 – 9) as even or odd.

Procedure. Before beginning the experiment, participants were told that they would read four passages for a memory test two days later. In phase 1, participants were asked to read the first prose passage at their own pace. All passages were presented on paper. After finishing, the passage was taken away and phase 2 began.

In Phase 2, the seven key concept sentences related to the just-read passage were presented for additional learning. Each sentence was presented on the computer one at a time. In the retrieval condition, participants were presented with the fill-in-the-blank versions and instructed to provide the answer for each. Specifically, in the FA-retrieval condition, participants were told to read the first part of the sentence silently and then provide the correct response vocally. They were told that this would help them learn the material from the passage in preparation of the later test. Before beginning, they were also instructed to provide their response within the 12 second time limit. After 12 seconds and regardless of

their response, the correct answer was displayed as feedback for 6 seconds. Participants were told to read the correct response out loud if it did not match their response or if no response was given. A blank screen then appeared for 1 second. The next fill-in-the-blank sentence was shown immediately after the blank screen and the process was repeated for all seven questions.

In the FA-restudy condition, participants saw the same key concept used in the retrieval condition, but with the correct response in place of the blank. Participants were told that this would provide an additional opportunity to learn material from the passage. They had 18 seconds to read the sentence silently and study it for the later memory test. As before, a blank screen appeared afterwards for 1 second before the next sentence was displayed, until all seven concepts had been presented.

The DA-restudy and DA-retrieval conditions followed the same procedure, with the addition of the secondary task. The task was similar to the digit-task used in Experiment 1, but with timing-modifications to reflect the longer phase 2 time limits and the more complex memory task. Specifically, digits (1-9) were played at a rate of 1 every 2 seconds for a total of 9 digits per sentence, starting with the onset of the sentence. As in Experiment 1, participants were told to use the keyboard to indicate if the digit was odd (i.e. press the “o” button) or even (i.e. press the “e” button). Again, buzzer feedback was used for incorrect or late responses. The task was continually performed throughout the entire set of seven sentences.

After completing phase 2 for the first passage, participants began phase 1 for second passage. Phases 1 and 2 were repeated for all four passages. Critically, phase 2 was

counterbalanced across participants such that each passage was equally distributed among the four possible conditions (i.e. FA-restudy, FA-retrieval, DA-restudy, and DA-retrieval).

After completing the first session, participants left the lab and were instructed to return two days later for the final test. Upon returning, participants took the final test (phase 3), consisting of the same four sets of seven questions each used during phase 2, now written on a sheet of paper. The sentences were all in fill-in-the-blank format and presented under FA. Participants were told that they had 12 minutes to answer all of the questions by writing the correct response in the blank space.

Scoring. Responses on the fill-in-the-blank questions were scored using two different methods, adapted from the free-response scale used by Agarwal et al. (2008). Because the current study used more constrained fill-in-the-blank questions, the scales did not cover as wide of a range as the scale Agarwal et al. (2008) used (i.e. 4 possible scoring options). Specifically, the all-or-nothing scale gave 1 point for correct and complete answers and 0 points for incorrect, incomplete, or missing answers (i.e. 2 possible scoring options). However, because many of the answers contained multiple words, another scale was used that allowed for more lenient grading. Specifically, the partial-credit scale gave 2 points for correct and complete answers, 1 point for correct, but not fully complete answers, and 0 points for incorrect, very incomplete, or missing answers (i.e. 3 possible scoring options). For example, if a participant gave the answer “sand bubblers” to the question described above, their score would be 0 for the all-or-nothing scale and 1 for the partial-credit scale. Likewise, if their response were “sand boils”, their score would be 1 for the all-or-nothing scale and 2 for the partial-credit scale. Finally, and similar to Agarwal et al. (2008), two raters independently scored the tests and the Pearson product moment correlation between their

scores was calculated after a subset of tests (i.e. the first 13 subjects). Because the inter-rater reliability was higher than .9, the acceptable cutoff value used in Agarwal et al. (2008), one rater scored the remaining tests. Specifically, the inter-rater reliability values were 1.00 for the all-or-nothing scale and 0.96 for the partial-credit scale.

Results

On average, participants spent 263.7 s reading a single passage during phase 1. The average time (s) spent reading each specific passage was roughly equivalent (i.e. Earthquakes: M = 258.88, SD = 58.08; Fossils: M = 261.6, SD = 62.25; Voyagers: M = 277.25, SD = 68.05; Wolves: M = 257.08, SD = 58.33).

All phase 2 (i.e. initial) and phase 3 (i.e. final) recall results were scored twice, once using the all-or-nothing scale and once using the partial-credit scale. Although this resulted in different descriptive statistics, all of the inferential tests produced identical conclusions. Therefore, the results using the all-or-nothing scale are reported in Appendix B and the results using the partial-credit scale are reported below.

Phase 2 Cued-Recall. The proportion of correctly answered fill-in-the-blank questions during phase 2 was M = .52 (SD = .24) and M = .49 (SD = .23), in the FA and DA conditions, respectively. Dividing attention during phase 2 did not significantly reduce phase 2 recall, $t(39) = .489$, $p = .628$.

Phase 3 Cued-Recall. On average, it took participants 440.13 s (SD = 144.84 s) to complete the final test. The phase 3 recall scores (Figure 5) were analyzed with a 2 x 2 repeated measures ANOVA, with phase 2 attention (FA vs. DA) and phase 2 condition (restudy vs. retrieval) as within-subjects factors. The results were similar to those obtained in Experiment's 1 and 2, specifically: (1) the main effect of attention was significant, $F(1, 39) =$

17.821, MS_e = .029, p < .001, η^2_p = .314, with higher recall in the FA than DA condition; (2) the main effect of condition was significant, $F(1, 39) = 16.44$, MS_e = .028, p < .001, η^2_p = .297, with retrieval practice producing greater final recall than restudy (i.e., a testing effect); and (3) the interaction between attention and condition was non-significant, $F(1, 39) = .354$, p = .555.

Phase 2 Digit Classification Task. Accuracy on the digit classification task was compared between the restudy and retrieval groups (Table 2). When assessing all trials, the restudy condition had significantly higher accuracy on the secondary task than the retrieval condition, $t(39) = 9.117$, SE = .051, p < .001, d = 2.039. This pattern remained when restricted to only those trials with actual responses, $t(39) = 6.043$, SE = .008, p < .001, d = 1.131.

Discussion

In general, the pattern of results in Experiment 3, using fill-in-the-blank, short answer questions, was similar to the pattern of results in Experiment's 1 and 2, which used Swahili-English word pairs. First, a benefit of retrieval practice compared to restudy emerged under both attention conditions. Second, phase 2 DA similarly impaired final recall in the restudy and retrieval conditions. Third, the interaction between phase 2 attention and condition was non-significant, as in Experiment's 1 and 2.

Experiment 3 provides additional evidence against the predictions of the effortful and elaborative accounts of the testing effect. Specifically, DA during phase 2 retrieval practice impaired the encoding effects of retrieval to a similar degree as restudy, replicating the results of Experiment's 1 and 2. However, the testing effect also did not increase in size under DA compared to FA, as it did in prior research (Buchin & Mulligan, 2017; Mulligan &

Picklesimer, 2016). As with Experiment's 1 and 2, this important difference may reflect the change in memory task materials (I return to this issue in the General Discussion).

One difference between the results of Experiment 3 and Experiment's 1 and 2 is that DA during phase 2 did not significantly disrupt phase 2 recall as it did in the prior experiments. This may be due to the change in the memory task materials. Specifically, the current task is easier in several ways than the task in Experiment's 1 and 2 – in particular, it does not require novel language learning, it provides fairly complete retrieval cues, and the syntactic structures of the sentences may scaffold learning. However, the resilience of retrieval under DA found in Experiment 3 does replicate prior research on the differential effects of DA on encoding and retrieval (e.g. Craik et al., 1996). Finally, the digit classification task results replicate the pattern in Experiments 1 and 2 as well as support prior research on the obligatory and attention-demanding nature of retrieval (e.g., Anderson et al., 1998; Craik et al., 1996; Naveh-Benjamin et al., 2000). Specifically, secondary task accuracy was higher when paired with restudy, than when paired with retrieval.

CHAPTER 5: GENERAL DISCUSSION

The primary goal of the current study was to assess the impact of DA on the encoding effects of retrieval while using materials more similar to those used in real-world learning. Prior research on DA and the encoding effects of retrieval support the idea that the encoding effects of retrieval are largely resilient to distraction, like retrieval success itself (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). Specifically, the prior research found that the testing effect actually increased in size under DA compared to FA. Alternatively stated, DA disrupted the encoding effects of restudy to a much greater degree than the encoding effects of retrieval. These prior findings conflict with predictions of testing effect accounts that hinge on effortful or elaborative processing, which are both disrupted by distraction (e.g., Craik & Broadbent, 1983; Craik et al., 1996; Hasher & Zacks, 1979; McDowd & Craik, 1988; Mulligan, 2008). Specifically, if the mnemonic benefits of retrieval did depend on greater effortful or elaborative processing than restudy, DA during phase 2 should have greatly reduced the size of the testing effect, although this was not the case.

Although prior research does not support these predictions, both studies (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016) used materials and methods typical of basic research in memory and are therefore limited in ecological validity and educational application. Because a chief motivation for research on the testing effect is its applicability to real-world learning, it is important to assess the impact of DA on the encoding effects of retrieval using materials more typical of educational settings. Perhaps the resilient benefit of retrieval practice would not extend to more complex, educationally relevant materials. If a

major goal behind studying the testing effect is to convince educators and students of the mnemonic benefits of testing, even with distraction, then it is important to do so using materials similar to those used in the classroom.

The critical results of the current study provide no support for the predictions of the effortful and elaborative accounts of the testing effect but also do not fully replicate the pattern found in the prior research. Specifically, across all 3 experiments the amount of disruption caused by DA during phase 2 was similar when paired with either restudy or retrieval. This similarity is not likely due to low power, as the current experiments each tested 40 participants and the power analysis discussed in Experiment 1 found that 24 participants were required to observe a significant interaction at .95 power (Faul et al., 2007). This in turn means that the size of the testing effect in the DA condition was similar to the size in the FA condition. However, prior research generally found that the testing effect increased in size under DA, compared to FA (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). Importantly, this may due to the educationally relevant materials used in the current study, which have generally been characterized as both difficult and complex (Carpenter et al., 2006; Grimaldi & Karpicke, 2014; Ozubko, Houriban, & MacLeod, 2012; Richland, Bjork, Finley, & Linn, 2005). In fact, although prior research using complex educationally relevant materials has typically found testing effects (e.g. Butler & Roediger, 2007; McDaniel, Anderson et al., 2007; Roediger & Karpicke, 2006), some studies argue against these findings (e.g. de Jonge, Tabbers, & Rikers, 2015; Van Gog & Sweller, 2015; but see Karpicke & Aue, 2014). Additionally, the deleterious effects of DA tend to increase along with task difficulty (e.g. McDowd & Craik, 1988). Overall, this discrepancy in results reflects the importance of the memory task materials and ecological validity as a whole.

The phase 2 recall and digit classification task results are also generally in line with prior research but some discrepancies warrant additional discussion. First, in Experiments 1 and 2, but not 3, DA during phase 2 significantly reduced initial recall compared to FA. Although DA during retrieval has been shown to be much less disruptive than DA during study, it can still produce modest, sometimes significant, reductions in recall success (e.g., Craik et al., 1996). Two methodological differences between the current experiments may explain the difference in DA disruption. First, the digit classification task in Experiment 3 played digits every 2 s, whereas the rate was faster (1.5 s) in Experiment's 1 and 2. Second, Experiment's 1 and 2 required participants to form novel associations (i.e. Swahili-English word pairs) using novel items (i.e. Swahili words), whereas Experiment 3 required participants to form novel associations using known words. Perhaps the former was more attention demanding than the latter. Additionally, the detailed and specific cues in Experiment 3 may have constrained retrieval more so than a novel Swahili word. Prior research indicates that retrieval success on tests with greater retrieval support is less likely to be disrupted by DA (e.g. Anderson et al., 2000; Craik & Broadbent, 1983; Craik et al., 1996; Craik & McDowd, 1987; Naveh-Benjamin, Craik, Guez, & Kreuger, 2005), making this difference potentially important. Ultimately, the different pattern of phase 2 recall was likely due to a combination of the slower DA task, the novelty of the Swahili words, and the enhanced retrieval support from the fill-in-the-blank short answer questions.

Second, all three experiments demonstrated the typical pattern of greater secondary task costs when paired with retrieval than restudy. This supports the results of Craik et al. (1996) as well as their analysis of the obligatory and attention-demanding nature of retrieval (e.g. Anderson et al., 1998; Naveh-Benjamin et al., 2000). Although this was expected in

Experiments 1 and 3, Experiment 2 was designed to address the potential ambiguity due to differences in secondary task performance between restudy and retrieval. The modifications to the secondary task in Experiment 2 were sufficient in better equating performance between the two conditions, however retrieval continued to significantly disrupt performance on the digit classification task more so than restudy. As previously discussed, these modifications were chosen instead of requiring a certain level of performance on the secondary task in order to maintain ecological validity. Therefore, while performance on the digit classification task did differ between conditions, the difference was numerically smaller than in Experiment 1 and the results reflect how participants naturally coordinate multiple cognitive demands.

Two related methodological aspects of Experiment 3 are also important to discuss. First, in phase 2 of Experiment's 1 and 2, participants were asked to read the target word out loud while restudying. This vocal response was necessary to equate the overt response in the restudy and retrieval conditions. In the phase 2 restudy condition of Experiment 3 however, participants were instructed to simply read and study the key concept sentence silently. Second, the target during phase 2 restudy was also more clearly defined in Experiment's 1 and 2 than in Experiment 3. Specifically, the key concepts were presented for restudy as normal sentences without the target words underlined or highlighted. Although participants did know the final test would consist of fill-in-the-blank questions (with the blank at the end of the sentence), the specific target words were not as obvious as they were in the retrieval condition.

The reasoning behind these methodological decisions was twofold. First, asking participants to alternate between silently reading most of a sentence before vocally reading

the remaining few words is not very ecologically valid in terms of typical study and review behavior. Second, I wished to keep the current procedures consistent with prior research to enhance its comparability. Prior testing-effect research using educationally valid materials typically did not require participants to provide a vocal or written response while restudying as they did while retrieving (Agarwal et al., 2008; Butler & Roediger, 2007; Kang et al., 2007; McDaniel, Anderson et al., 2007). Similarly, participants in the restudy condition were not explicitly told or shown the specific words or concepts that would be targets on the final test (Agarwal et al., 2008; Butler & Roediger, 2007; Kang et al., 2007; McDaniel, Anderson et al., 2007).

That said, it should be acknowledged that the design of typical testing-effect experiments with educationally valid materials, as well as the design of the current Experiment 3, may advantage the retrieval condition. Not only are the exact targets not highlighted in the restudy condition, but participants are typically not required to produce a written or vocal response as they are in the retrieval condition. Highlighting the exact target items could plausibly enhance attention to that part of the material, affecting the extent to which it is rehearsed. Thus, the retrieval condition could benefit from that knowledge over-and-above any effect of retrieval itself. Another advantage concerns the production effect – the robust finding that producing a word aloud during study, relative to simply reading a word silently, improves later memory (e.g. MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010; Ozubko et al., 2012). Research on the educational benefits of retrieval practice would benefit from fully equating target emphasis and production in the retrieval and restudy conditions. Importantly, because a general goal of this research is to enhance its educational relevance, the conditions should be equated without sacrificing ecological validity. Along

those same lines, the applicability of research on attention and the testing effect would profit from assessing the effects of passive distraction on the encoding effects of restudy and retrieval. Real-world learning often requires students to overcome passive distraction, like when reading in a noisy library or studying while listening to music or TV.

TABLES

Table 1.

Phase 2 Cued-Recall Proportion Correct: Mean (SD)

	Full Attention (FA)	Divided Attention (DA)
Experiment 1		
1-rep	.13 (.16)	.08 (.13)
3-rep	.38 (.18)	.22 (.17)
Experiment 2		
1-rep	.11 (.13)	.13 (.13)
3-rep	.35 (.17)	.20 (.16)
Experiment 3	.52 (.24)	.49 (.23)

Note. The recall results of Experiment 3 were scored using the partial-credit scale.

Table 2.

Phase 2 Digit Classification Task Proportion Correct and Reaction Time: Mean (SD)

	All trials (includes omissions)		Only responses (excludes omissions)	
	Restudy	Retrieval	Restudy	Retrieval
Experiment 1				
1-rep	.85 (.12)	.66 (.15)	.92 (.07)	.85 (.10)
3-rep	.85 (.10)	.65 (.12)	.93 (.05)	.86 (.07)
Experiment 2				
1-rep	.92 (.14)	.80 (.18)		
3-rep	.94 (.07)	.79 (.14)		
Experiment 3	.96 (.06)	.77 (.12)	.98 (.03)	.93 (.05)

FIGURES

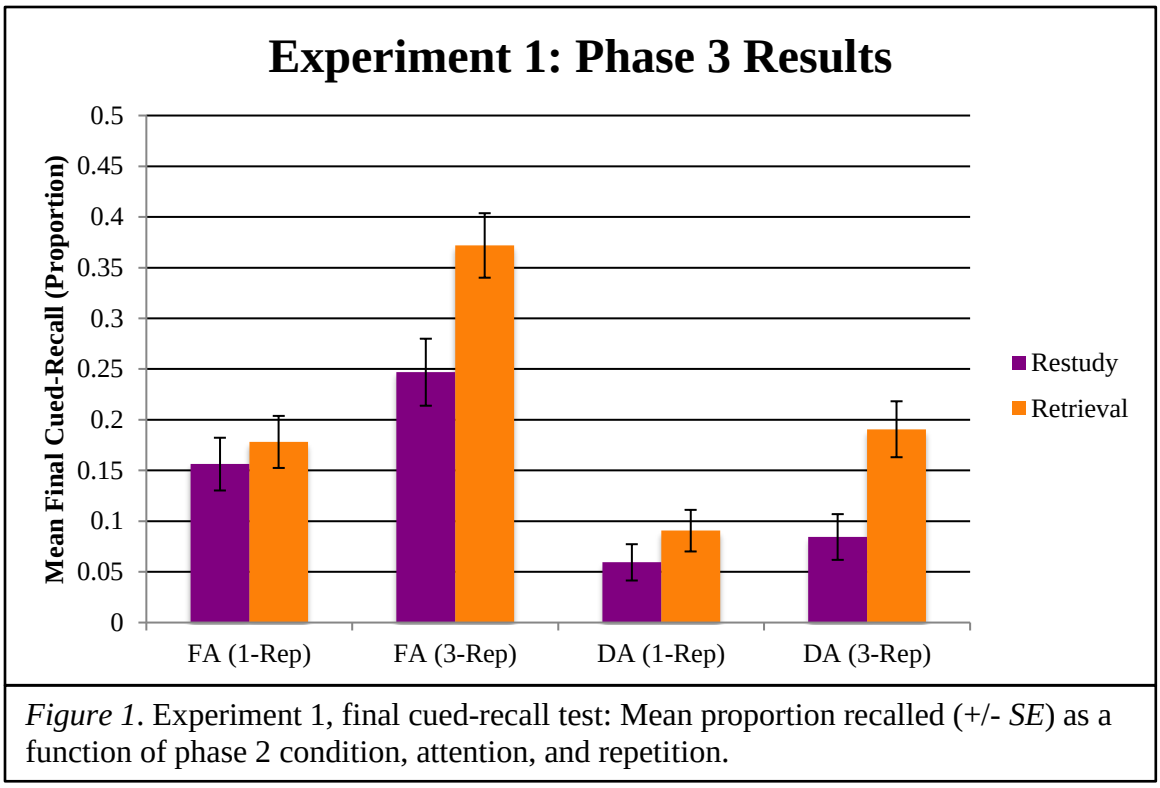

Figure 1. Experiment 1, final cued-recall test: Mean proportion recalled (+/- *SE*) as a function of phase 2 condition, attention, and repetition.

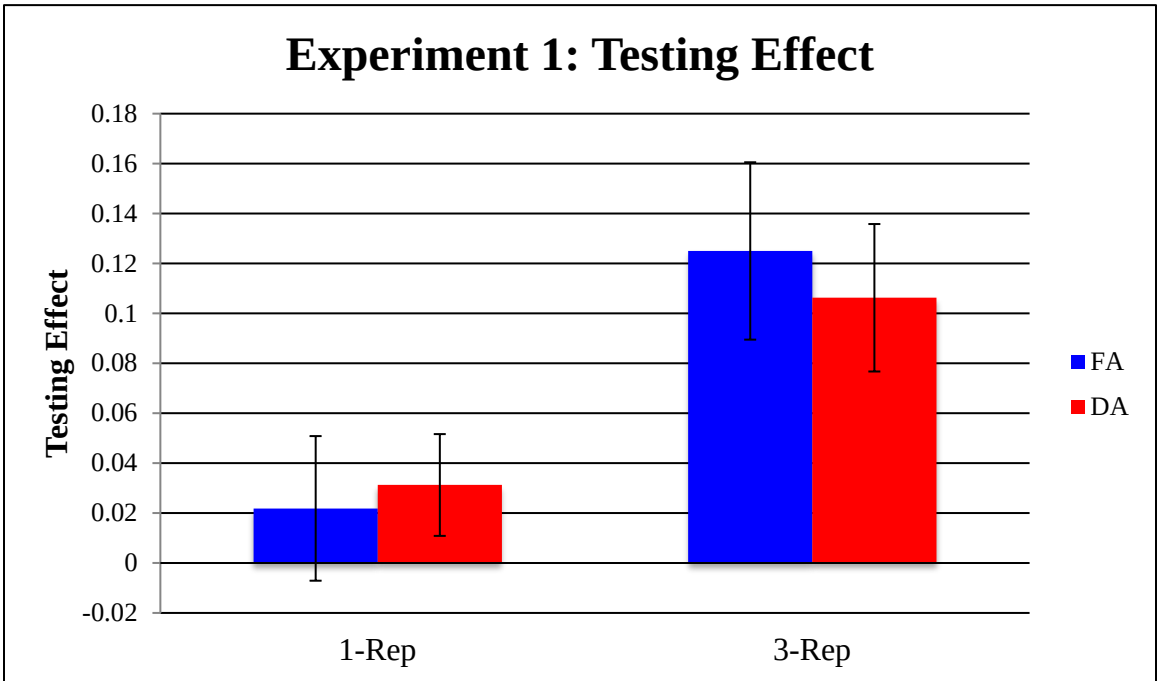

Figure 2. Experiment 1, final cued-recall test: Mean testing effect (+/- *SE*) as a function of phase 2 attention and repetition.

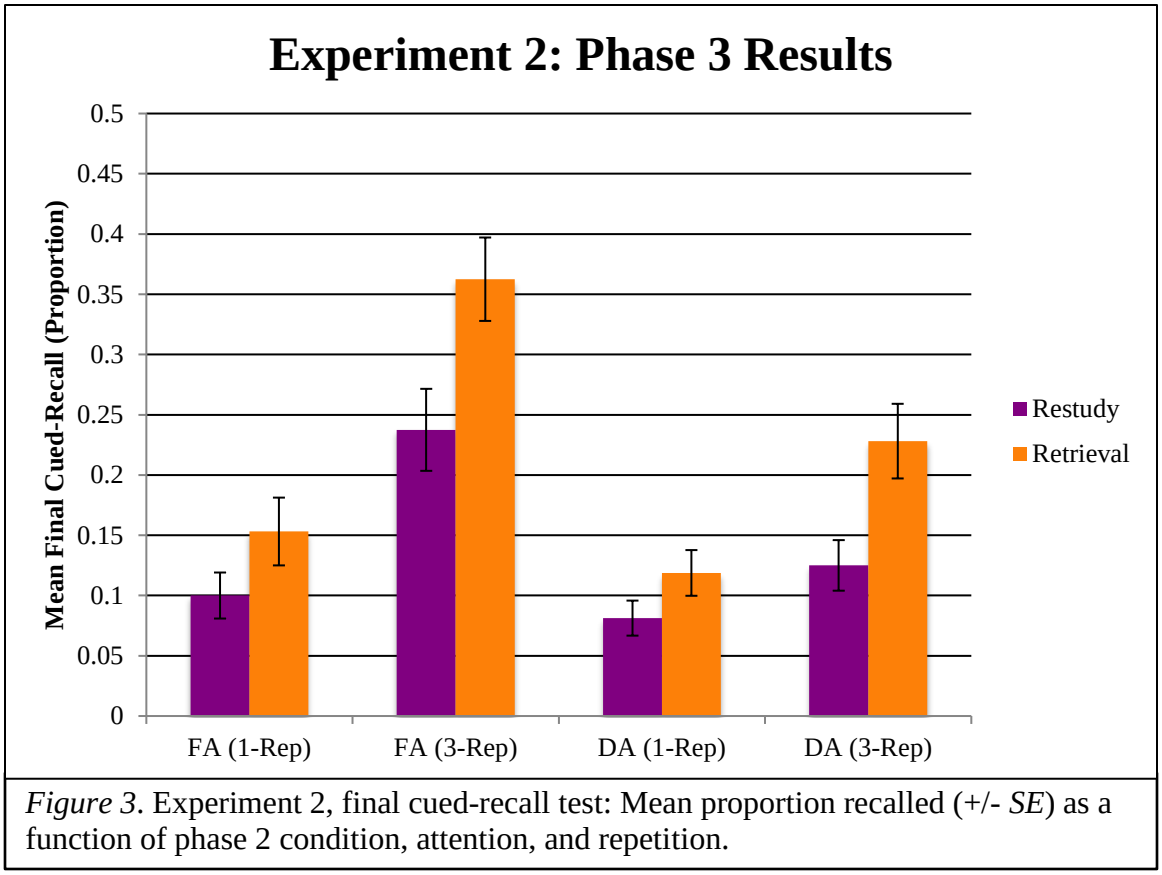

Figure 3. Experiment 2, final cued-recall test: Mean proportion recalled (+/- *SE*) as a function of phase 2 condition, attention, and repetition.

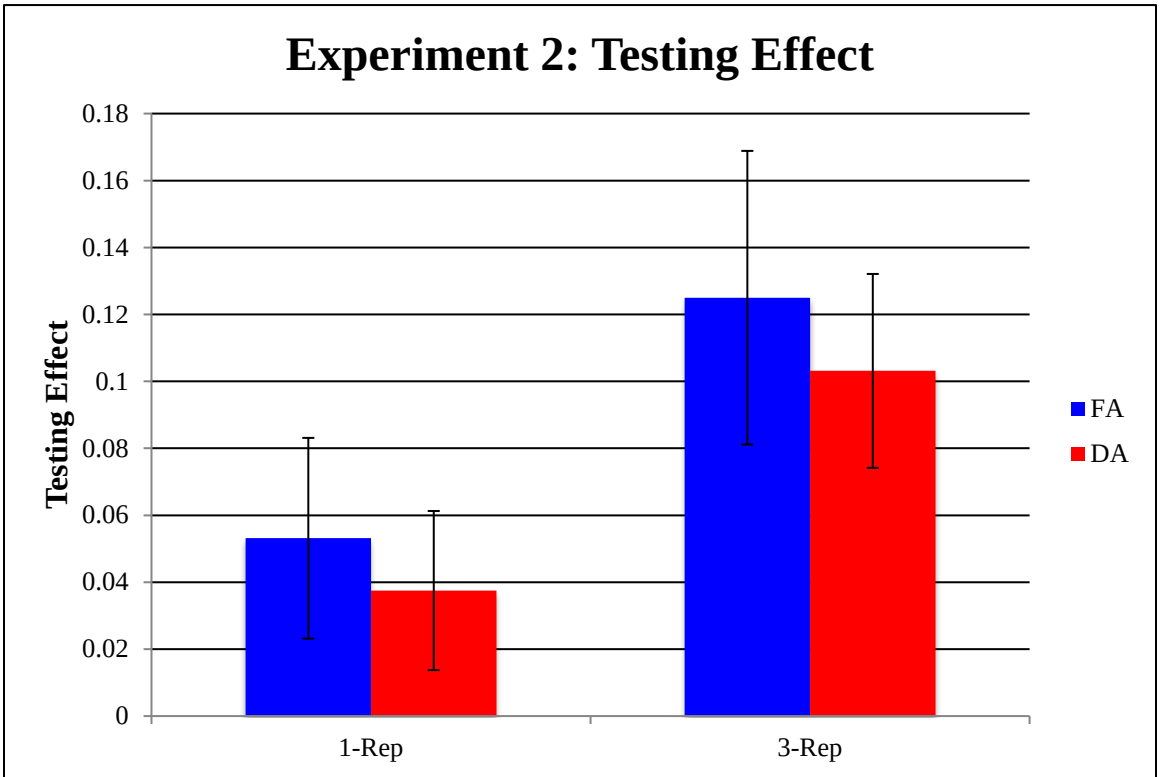

Figure 4. Experiment 2, final cued-recall test: Mean testing effect (+/- *SE*) as a function of phase 2 attention and repetition.

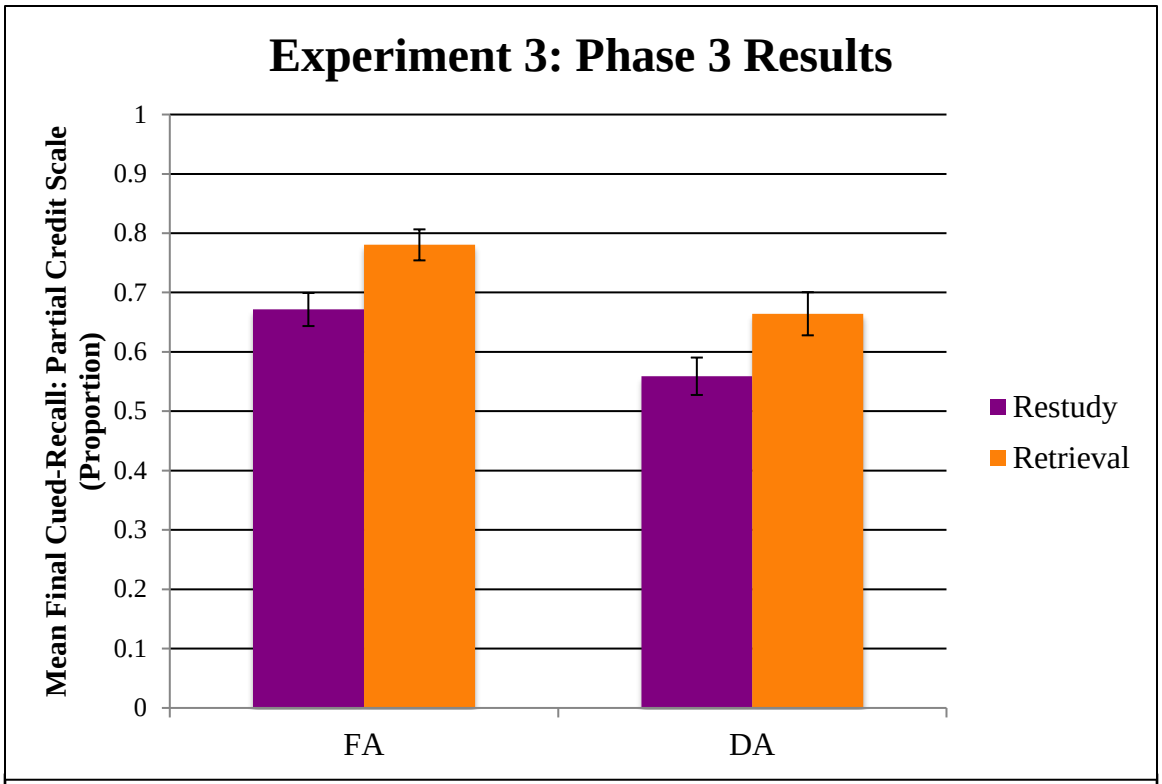

Figure 5. Experiment 3, final cued-recall fill-in-the-blank test: Mean proportion recalled (+/- *SE*) as a function of phase 2 condition and attention.

APPENDIX A: CONDITIONALIZED ANALYSES

Testing-effect research often includes a conditionalized recall measure. The conditionalized recall score was calculated as the proportion of target words recalled out of only those words successfully retrieved at least once during phase 2. The conditionalized analyses of all 3 experiments are outlined below. The advantages and disadvantages of the unconditionalized and conditionalized measures will be briefly discussed first. The unconditionalized recall score, although a widely used and straightforward measure, produces a re-experience confound (e.g. Kuo & Hirshman, 1996; Rowland & DeLosh, 2015; Toppino & Cohen, 2009). Specifically, in the restudy condition, all words are restudied and re-experienced during phase 2 but, in the retrieval condition, words that were actually retrieved are mixed together with words that were not retrieved and merely re-experienced through feedback. This may advantage the restudy condition and underestimate the mnemonic benefits of retrieval practice. Additionally, this problem is amplified when comparing multiple retrieval conditions, which differ in phase 2 success rate, as in the current study.

Conditionalizing final recall on successful phase 2 recall eliminates the re-experience confound. Specifically, the restudy condition is unaffected and continues to include only those words re-experienced through restudy during phase 2, whereas the retrieval condition now includes only those words re-experienced through successful recall during phase 2. In other words, the retrieval condition no longer includes words that were merely re-experienced through feedback. Because I am primarily concerned with the encoding effects of retrieval, it seems reasonable to focus on only those items that were actually retrieved during phase 2. Despite these advantages, the conditionalized measure raises a different

concern – the possibility of item selection effects. Using this measure results in different subsets of target items in the restudy and retrieval conditions (and across other retrieval conditions that may differ in phase 2 retrieval rate). The general low rate of phase 2 recall in the current experiments amplifies this concern. Therefore, the conditionalized recall results are reported with limited interpretations below.

Experiment 1

The conditionalized recall scores (Table A1) were submitted to a 2 (phase 2 attention) x 2 (phase 2 condition) x 2 (phase 2 repetition) ANOVA, which revealed the following results:

- A significant main effect of phase 2 attention, $F(1, 39) = 21.305$, $MS_e = .069$, $p < .001$, $\eta^2_p = .353$, with lower final recall for phase 2 DA, than FA.
- A significant main effect of phase 2 condition, $F(1, 39) = 84.446$, $MS_e = .054$, $p < .001$, $\eta^2_p = .684$, indicating a benefit of phase 2 retrieval practice compared to restudy (i.e. a testing effect).
- A significant main effect of phase 2 repetition, $F(1, 39) = 26.175$, $MS_e = .061$, $p < .001$, $\eta^2_p = .402$, with higher final recall for 3-rep than 1-rep.
- A non-significant two-way interaction between phase 2 attention and condition, $F(1, 39) = 0.034$, $p = .855$.
- A non-significant two-way interaction between phase 2 attention and repetition, $F(1, 39) = 0.294$, $p = .591$.
- A significant two-way interaction between phase 2 condition and repetition, $F(1, 39) = 9.046$, $MS_e = .061$, $p = .005$, $\eta^2_p = .188$.

- A non-significant three-way interaction between phase 2 attention, condition, and repetition, $F(1, 39) = 0.482$, $p = .492$.

The significant interaction between phase 2 condition and repetition indicates that the testing effect was larger in the 3-rep, than 1-rep, condition. Because the interaction was significant, additional tests are needed to see if the benefit of retrieval practice is significant when only assessing 1-rep pairs. Therefore, the benefit of retrieval practice was assessed for words repeatedly experienced during phase 2 as well as for words only experienced once during phase 2. A 2 (attention) x 2 (phase 2 condition) ANOVA was conducted at each level of phase 2 repetition (1-rep vs. 3-rep). The two 2 x 2 ANOVA's, for the phase 2 1-rep condition and the phase 2 3-rep condition, respectively, revealed the following:

- A significant main effect of phase 2 attention, $F(1, 39) = 6.824$, $MS_e = .086$, $p = .013$, $\eta^2_p = .149$, indicating lower final recall for phase 2 DA, than FA.
- A significant main effect of phase 2 condition, $F(1, 39) = 18.672$, $MS_e = .051$, $p < .001$, $\eta^2_p = .324$, demonstrating a testing effect.
- A non-significant two-way interaction between phase 2 attention and condition, $F(1, 39) = 0.253$, $p = .618$.
- A significant main effect of phase 2 attention, $F(1, 39) = 22.3$, $MS_e = .041$, $p < .001$, $\eta^2_p = .364$, with lower final recall for phase 2 DA, than FA.
- A significant main effect of phase 2 condition, $F(1, 39) = 64.843$, $MS_e = .064$, $p < .001$, $\eta^2_p = .624$, revealing a testing effect.
- A non-significant two-way interaction between phase 2 attention and condition, $F(1, 39) = 0.111$, $p = .741$.

The results indicate that the beneficial effect of retrieval practice, compared to restudy (i.e. the testing effect), was significant in both phase 2 repetition conditions, although the effect was much larger in the 3-rep, than 1-rep, condition. Alternatively interpreted, the two-way interaction between phase 2 condition and repetition also indicates that the benefit of phase 2 repetition was larger for phase 2 retrieval, than restudy. To determine if repetition during phase 2 was beneficial in the restudy condition, a 2 (attention) x 2 (repetition) ANOVA was conducted at each level of phase 2 condition (restudy vs. retrieval). The two 2 x 2 ANOVA's, for the phase 2 restudy condition and the phase 2 retrieval condition, respectively, revealed the following:

- A significant main effect of phase 2 attention, $F(1, 39) = 29.105$, $MS_e = .023$, $p < .001$, $\eta^2_p = .427$, with lower final recall for phase 2 DA, than FA.
- A significant main effect of phase 2 repetition, $F(1, 39) = 6.273$, $MS_e = .021$, $p = .017$, $\eta^2_p = .139$, with higher final recall for words repeated during phase 2.
- A non-significant two-way interaction between phase 2 attention and repetition, $F(1, 39) = 3.162$, $p = .083$
- A significant main effect of phase 2 attention, $F(1, 39) = 5.857$, $MS_e = .138$, $p = .02$, $\eta^2_p = .131$, with lower final recall for phase 2 DA, than FA.
- A significant main effect of phase 2 repetition, $F(1, 39) = 19.97$, $MS_e = .101$, $p < .001$, $\eta^2_p = .339$, with higher final recall for words repeated during phase 2.
- A non-significant two-way interaction between phase 2 attention and repetition, $F(1, 39) = 0.006$, $p = .941$.

The results indicate that the effect of phase 2 repetition was significant for both phase 2 conditions, although the repetition benefit was larger for the retrieval, than restudy,

condition. Taken together, the additional ANOVA results indicate that although there was a significant interaction between phase 2 condition and repetition, it did not stem from any non-significant main effects at specific levels of phase 2 condition or repetition.

Experiment 2

The conditionalized recall scores (Table A1) were submitted to the same 2 (phase 2 attention) x 2 (phase 2 condition) x 2 (phase 2 repetition) ANOVA, which revealed the following results:

- A non-significant main effect of phase 2 attention, $F(1, 39) = 1.879$, $p = .178$.
- A significant main effect of phase 2 condition, $F(1, 39) = 84.969$, $MS_e = .092$, $p < .001$, $\eta^2_p = .685$, indicating a benefit of phase 2 retrieval practice compared to restudy (i.e. a testing effect).
- A significant main effect of phase 2 repetition, $F(1, 39) = 18.88$, $MS_e = .061$, $p < .001$, $\eta^2_p = .326$, with higher final recall for 3-rep than 1-rep.
- A non-significant two-way interaction between phase 2 attention and condition, $F(1, 39) = 0.899$, $p = .349$.
- A marginally-significant two-way interaction between phase 2 attention and repetition, $F(1, 39) = 3.471$, $MS_e = .10$, $p = .07$, $\eta^2_p = .082$.
- A non-significant two-way interaction between phase 2 condition and repetition, $F(1, 39) = 0.763$, $p = .388$.
- A non-significant three-way interaction between phase 2 attention, condition, and repetition, $F(1, 39) = 0.299$, $p = .587$.

The marginally significant two-way interaction (i.e. $p = .07$) between phase 2 attention and repetition suggests that the negative effect of DA was greater in the 3-rep, than

1-rep, condition. Similar to the prior experiment, additional tests are needed to determine the significance of the phase 2 attention effect in the 1-rep condition. A 2 (attention) x 2 (phase 2 condition) ANOVA was conducted at each level of phase 2 repetition (1-rep vs. 3-rep). The two 2 x 2 ANOVA's, for the phase 2 1-rep condition and the phase 2 3-rep condition, respectively, revealed the following:

- A non-significant main effect of phase 2 attention, $F(1, 39) = 0.268$, $p = .608$.
- A significant main effect of phase 2 condition, $F(1, 39) = 26.185$, $MS_e = .123$, $p < .001$, $\eta^2_p = .402$, indicating a testing effect.
- A non-significant two-way interaction between phase 2 attention and condition, $F(1, 39) = 0.818$, $p = .371$.
- A significant main effect of phase 2 attention, $F(1, 39) = 8.0$, $MS_e = .055$, $p = .007$, $\eta^2_p = .17$, with lower final recall for phase 2 DA, than FA.
- A significant main effect of phase 2 condition, $F(1, 39) = 77.147$, $MS_e = .061$, $p < .001$, $\eta^2_p = .664$, indicating a testing effect.
- A non-significant two-way interaction between phase 2 attention and condition, $F(1, 39) = 0.041$, $p = .84$.

The results indicate that the effect of phase 2 attention was significant for words repeated during phase 2, but not for words presented only once during phase 2. Alternatively, the marginally-significant interaction between phase 2 attention and condition can also be interpreted as a larger benefit of phase 2 repetition when under FA, than under DA. To determine whether repetition under DA significantly benefited later memory, a 2 (phase 2 condition) x 2 (repetition) ANOVA was conducted at each level of phase 2 attention

(FA vs. DA). The two 2 x 2 ANOVA's, for the phase 2 FA condition and the phase 2 DA condition, respectively, revealed the following:

- A significant main effect of phase 2 condition, $F(1, 39) = 47.787$, $MS_e = .068$, $p < .001$, $\eta^2_p = .551$, revealing a testing effect.
- A significant main effect of phase 2 repetition, $F(1, 39) = 23.165$, $MS_e = .06$, $p < .001$, $\eta^2_p = .373$, with higher final recall for words repeated during phase 2.
- A non-significant two-way interaction between phase 2 condition and repetition, $F(1, 39) = 1.031$, $p = .316$.
- A significant main effect of phase 2 condition, $F(1, 39) = 52.751$, $MS_e = .087$, $p < .001$, $\eta^2_p = .575$, indicating a testing effect.
- A non-significant main effect of phase 2 repetition, $F(1, 39) = 1.156$, $p = .289$.
- A non-significant two-way interaction between phase 2 condition and repetition, $F(1, 39) = 0.045$, $p = .833$.

The results reveal a similar pattern to the prior set of analyses, in that the effect of phase 2 repetition was significant under FA, but not under DA. Taken together, the additional ANOVA results reflect the significant interaction between phase 2 attention and repetition in that main effects were observed only at certain levels of attention and repetition. Specifically, DA during phase 2 disrupted later recall in the 3-rep, but not 1-rep, condition and repetition during phase 2 benefited later memory in the FA, but not DA, condition.

Experiment 3

The conditionalized fill-in-the-blank recall scores using the partial-credit scale (Table A1) were analyzed with a 2 x 2 ANOVA, with phase 2 attention (FA vs. DA) and phase 2 condition (restudy vs. retrieval) as within-subjects factors. The results of the repeated-

measures ANOVA, similar to Experiment 1 and, to a lesser extent, Experiment 2, revealed the following:

- A significant main effect of phase 2 attention, $F(1, 39) = 4.133$, $MS_e = .05$, $p = .049$, $\eta^2_p = .096$, indicating a benefit of phase 2 FA compared to DA.
- A significant main effect of phase 2 condition, $F(1, 39) = 129.773$, $MS_e = .029$, $p < .001$, $\eta^2_p = .769$, demonstrating a benefit of retrieval practice compared to restudy
- A non-significant interaction between phase 2 attention and condition, $F(1, 39) = 2.642$, $p = .112$.

Although the interaction was non-significant, a set of a-priori planned contrasts was conducted, revealing the following:

- A significant effect of phase 2 attention in the restudy condition, $t(39) = 3.405$, $SE = .033$, $p = .002$, $d = .598$, indicating lower final recall when restudying was done under DA, compared to under FA.
- A non-significant effect of phase 2 attention in the retrieval condition, $t(39) = 0.617$, $p = .54$.
- A significant effect of phase 2 condition in the FA condition, $t(39) = -8.29$, $SE = .033$, $p < .001$, $d = 1.324$, revealing a testing effect.
- A significant effect of phase 2 condition in the DA condition, $t(39) = -8.758$, $SE = .04$, $p < .001$, $d = 1.571$, revealing a testing effect.

The planned comparisons indicate two interesting findings, the latter of which may depend on successful phase 2 recall. First, the size of the testing effect was very similar under

FA and under DA. Second, DA during phase 2 did not significantly reduce final recall in the retrieval condition as it did in the restudy condition.

Table A1.

Retrieval Condition: Phase 3 Final Cued-Recall Proportion Correct; Conditionalized on Correct Recall in Phase 2: Mean (SD)

	Full Attention	Divided Attention
Experiment 1		
1-rep Condition	0.34 (0.43)	0.19 (0.38)
3-rep Condition	0.56 (0.24)	0.42 (0.35)
Experiment 2		
1-rep Condition	0.34 (0.46)	0.41 (0.48)
3-rep Condition	0.57 (0.24)	0.48 (0.37)
Experiment 3	0.98 (.22)	0.91 (.24)

Note. The recall results of Experiment 3 were scored using the partial-credit scale.

APPENDIX B: EXPERIMENT 3: ALL-OR-NOTHING SCALE RESULTS

The recall results of Experiment 3 using the all-or-nothing scale are outlined below. The results of all inferential tests, in terms of significant effects and differences, were identical to the results using the partial-credit scale and only differed in terms of descriptive statistics.

During phase 2, the proportion of correctly answered fill-in-the-blank questions was M = .44 (SD = .26) and M = .39 (SD = .23), in the FA and DA conditions, respectively. Dividing attention during phase 2 did not significantly reduce phase 2 recall, $t(39) = .929$, $p = .358$.

The unconditionalized phase 3 recall scores (Table B1) were analyzed with a 2 x 2 ANOVA, using phase 2 attention (FA vs. DA) and phase 2 condition (restudy vs. retrieval) as within-subjects factors. Both main effects, but not the interaction between attention and condition, were significant. The specific results are as follows:

- A significant main effect of phase 2 attention, $F(1, 39) = 14.34$, $MS_e = .041$, $p = .001$, $\eta^2_p = .269$.
- A significant main effect of phase 2 condition, $F(1, 39) = 18.91$, $MS_e = .031$, $p < .001$, $\eta^2_p = .327$.
- A non-significant interaction between phase 2 attention and condition, $F(1, 39) = .354$, $p = .555$.

Although the interaction was non-significant, the same set of a-priori planned contrasts carried out in the previous experiments was conducted using the all-or-nothing scale results. The planned contrasts revealed the following:

- A significant effect of phase 2 attention in the restudy condition, $t(39) = 2.354$, $SE = .044$, $p = .024$, $d = .475$.
- A significant effect of phase 2 attention in the retrieval condition, $t(39) = 3.176$, $SE = .044$, $p = .003$, $d = .614$.
- A significant effect of phase 2 condition in the FA condition, $t(39) = -3.267$, $SE = .043$, $p = .002$, $d = .659$.
- A significant effect of phase 2 condition in the DA condition, $t(39) = 2.354$, $SE = .039$, $p = .012$, $d = .445$.

Next, the conditionalized phase 3 recall scores (Table B1) were analyzed with the same 2 (attention) x 2 (phase 2 condition) ANOVA. As in the unconditionalized analysis, both main effects, but not the interaction between attention and condition, were significant. The specific results are as follows:

- A significant main effect of phase 2 attention, $F(1, 39) = 5.245$, $MS_e = .094$, $p = .028$, $\eta^2_p = .119$.
- A significant main effect of phase 2 condition, $F(1, 39) = 124.943$, $MS_e = .038$, $p < .001$, $\eta^2_p = .762$.
- A non-significant interaction between phase 2 attention and condition, $F(1, 39) = .043$, $p = .837$.

The same planned contrasts used with the unconditionalized data were used again with the conditionalized data, demonstrating the following (note that the restudy conditions are unaffected by this analysis):

- A non-significant effect of phase 2 attention in the retrieval condition, $t(39) = 1.608$, $p = .116$.

- A significant effect of phase 2 condition in the FA condition, $t(39) = -8.823.354$, $SE = .04$, $p < .001$, $d = 1.56$.
- A significant effect of phase 2 condition in the DA condition, $t(39) = -6.153$, $SE = .055$, $p < .001$, $d = 1.112$.

Table B1.

Phase 3 Final Cued-Recall Proportion Correct: Mean (SD)

	Unconditionalized		Conditionalized	
	Full Attention (FA)	Divided Attention (DA)	Full Attention (FA)	Divided Attention (DA)
Restudy	.58 (.22)	.47 (.22)		
Retrieval	.71 (.21)	.58 (.25)	.93 (.23)	.81 (.37)

Note. The recall results of Experiment 3 were scored using the all-or-nothing scale.

REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open-and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861-876.

Anderson, N. D., Craik, F. I., & Naveh-Benjamin, M. (1998). The attentional demands of encoding and retrieval in younger and older adults: I. Evidence from divided attention costs. *Psychology and Aging, 13*(3), 405.

Anderson, N. D., Iidaka, T., Cabeza, R., Kapur, S., McIntosh, A. R., & Craik, F. I. (2000). The effects of divided attention on encoding-and retrieval-related brain activity: A PET study of younger and older adults. *Journal of Cognitive Neuroscience, 12*(5), 775-792.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From Learning Processes To Cognitive Processes: Essays In Honor Of William K. Estes, 2*, 35-67.

Buchin, Z. L., & Mulligan, N. W. (2017). The testing effect under divided attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(12), 1934.

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

Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4-5), 514-527.

Calderwood, C., Ackerman, P. L., & Conklin, E. M. (2014). What else do college students "do" while studying? An investigation of multitasking. *Computers & Education, 75*, 19-29.

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

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

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448.

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633-642.

Chan, J. C., & 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.

Cooper, J. D., Pikulski, J. J., Au, K. H., Caldero´n, M., Comas, J. C., Lipson, M. Y., et al. (1996). *Explore*. Boston, MA: Houghton Mifflin Company.

Craik, F. I., & Broadbent, D. E. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *302*(1110), 341-359.

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2),159.

Craik, F. I., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 474.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*(3), 215-235.

de Jonge, M., Tabbers, H. K., & Rikers, R. M. (2015). The effect of testing on the retention of coherent and incoherent text material. *Educational Psychology Review*, *27*(2), 305-315.

Dudukovic, N. M., DuBrow, S., & Wagner, A. D. (2009). Attention during memory retrieval enhances future remembering. *Memory & Cognition*, *37*(7), 953-961.

Dudukovic, N. M., Gottshall, J. L., Cavanaugh, P. A., & Moody, C. T. (2015). Diminished testing benefits in young adults with attention-deficit hyperactivity disorder. *Memory*, *23*(8), 1264-1276.

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, *6*.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191.

Fernandes, M. A., & Moscovitch, M. (2000). Divided attention and memory: evidence of substantial interference effects at retrieval and encoding. *Journal of Experimental Psychology: General, 129*(2), 155.

Fernandes, M. A., & Moscovitch, M. (2002). Factors modulating the effect of divided attention during retrieval of words. *Memory & Cognition, 30*(5), 731-744.

Fernandes, M. A., & Moscovitch, M. (2003). Interference effects from divided attention during retrieval in younger and older adults. *Psychology and Aging, 18*(2), 219.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221.

Gaspelin, N., Ruthruff, E., & Pashler, H. (2013). Divided attention: An undesirable difficulty in memory retention. *Memory & Cognition, 41*(7), 978-988.

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology, 106*(1), 58.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*(3), 356.

Hicks, J. L., & Marsh, R. L. (2000). Toward specifying the attentional demands of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1483.

Jacobsen, W. C., & Forste, R. (2011). The wired generation: Academic and social outcomes of electronic media use among university students. *Cyberpsychology, Behavior, and Social Networking, 14*(5), 275-280.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621.

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558.

Kang, S. H., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation?. *Journal of Applied Research in Memory and Cognition, 3*(3), 183-188.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317-326.

Kessler, Y., Vandermorris, S., Gopie, N., Daros, A., Winocur, G., & Moscovitch, M. (2014). Divided attention improves delayed, but not immediate retrieval of a consolidated memory. *Plos One, 9*(3), e91309.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education, 43*(1), 21-27.

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology*, 451-464.

LeCompte, D. C., Neely, C. B., & Wilson, J. R. (1997). Irrelevant speech and irrelevant tones: The relative importance of speech to the irrelevant speech effect. J*ournal of Experimental Psychology: Learning, Memory, and Cognition, 23*(2), 472.

Lozito, J. P., & Mulligan, N. W. (2010). Exploring the role of attention during implicit memory retrieval. *Journal of Memory and Language*, *63*(3), 387-399.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671.

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

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200-206.

McDowd, J. M., & Craik, F. I. (1988). Effects of aging and task difficulty on divided attention performance. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 267.

Mulligan, N. W. (2008). Attention and memory. *Learning and Memory: A Comprehensive Reference*, *2*, 7-22.

Mulligan, N. W., & Lozito, J. P. (2006). An asymmetry between memory encoding and retrieval revelation, generation, and transfer-appropriate processing. *Psychological Science*, *17*(1), 7-11.

Mulligan, N. W., & Picklesimer, M. (2016). Attention and the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 938.

Murdock, B. B. (1965). Effects of a subsidiary task on short-term memory. *British Journal of Psychology, 56*(4), 413-419.

Naveh-Benjamin, M., Craik, F. I., Guez, J., & Kreuger, S. (2005). Divided attention in younger and older adults: effects of strategy and relatedness on memory performance and secondary task costs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 520.

Naveh-Benjamin, M., Craik, F. I., Perretta, J. G., & Tonev, S. T. (2000). The effects of divided attention on encoding and retrieval processes: The resiliency of retrieval processes. *The Quarterly Journal of Experimental Psychology: Section A, 53*(3), 609-625.

Naveh-Benjamin, M., Kilb, A., & Fisher, T. (2006). Concurrent task effects on memory encoding and retrieval: Further support for an asymmetry. *Memory & Cognition, 34*(1), 90-101.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2*(3), 325-335.

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory, 20*(7), 717-727.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory and Language, 60*(4), 437-447.

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

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test–restudy practice: Implications for student learning. *Applied Cognitive Psychology, 25*(1), 87-95.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition, 43*(4), 619-633.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1850-1855).

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27.

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

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

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory, 23*(3), 403-419.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2008). When attention matters: The curious incident of the wandering mind. *Memory & Cognition, 36*(6), 1144-1150.

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology, 36*, 1710–1727.

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, 110*(16), 6313-6317.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252-257.

van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Wirebring, L. K., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the "testing effect": A review. *Trends in Neuroscience and Education*.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247-264.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214-221.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, 20*(1), 6-10.

Yale University (2005). *The Kamusi project: Internet living Swahili dictionary*. Retrieved March 25, 2005, from www.kamusiproject.org.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*(8), 995-1008.