THESIS

TESTING EFFECTS IN CONTEXT MEMORY

Submitted by

Christopher A. Rowland

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2011

Master's Committee:

Advisor: Edward DeLosh

Matthew Rhodes
Charles Anderson

ABSTRACT

TESTING EFFECTS IN CONTEXT MEMORY

Retrieving a previously learned piece of information can have profound positive effects on the later retention of such information. However, it is not clear if test-induced memory benefits are restricted to the specific information which was retrieved, or if they can generalize more completely to the full study episode. Two experiments investigated the role of retrieval practice on memory for both target and non-target contextual information. Experiment 1 used a remember-know task to assess the subjective quality of memory as a function of earlier retrieval practice or study. Additionally, memory for context information (target font color) from the initial study episode was assessed. Experiment 2 used paired associates to investigate the effect of testing on non-tested but associated contextual information. Successful retrieval practice, compared with study, resulted in large benefits in target, target-associated, and context information retention across both experiments. Moreover, successful retrieval practice was associated with a greater contribution of remember responses informing recognition decisions. The results suggest that retrieving information may serve to both boost item memory about a target and strengthen the bind between target and associated contextual information. In sum, the present study adds to an emerging literature that test-induced mnemonic benefits may "spill over" to non-tested information.

DEDICATION

I dedicate this thesis to my grandfather, Alson F. Pierce.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Retrieval is one of the most important components in the study of human memory (Roediger, 2000). The effects of testing memory stand out as a clear testament to this notion. A robust finding in the memory literature is that when previously studied information is subject to a test, it is more likely to be later remembered than similar information not subject to a test (see Butler & Roediger, 2011; Roediger & Karpicke, 2006a, for a review). This finding, termed the testing effect, has been found under a wide variety of circumstances. Verbal materials are often found to be subject to testing effects, including single word lists (e.g., Carpenter & DeLosh, 2006; Kuo & Hirshman, 1996, 1997; Szpunar, McDermott, & Roediger, 2008; Zaromb & Roediger, 2010), paired associates (e.g., Allen, Mahler, & Estes, 1969; Carpenter, 2009; Carpenter, Pashler, & Vul, 2006; Carrier & Pashler, 1992; Pyc & Rawson, 2010; Runquist, 1983, 1986; Toppino & Cohen, 2009), and prose materials (e.g., Chan, McDermott, & Roediger, 2006; Glover, 1989; Hinze & Wiley, 2011; Karpicke & Blunt, 2011; Roediger & Karpicke, 2006b; Roediger & Marsh, 2005). In addition, the testing effect generalizes to verbal-nonverbal associative information (e.g., Carpenter & DeLosh, 2005; Tse, Balota, & Roediger, 2010), visuospatial information (e.g., Carpenter & Pashler, 2007; Kang, 2010; Rohrer, Taylor, & Sholar, 2010) and visual – spatial location associations (Sommer, Schoell, & Buchel, 2008). Outside the laboratory, testing has been successfully applied to improve performance on a variety of tasks in ecologically valid classroom settings (e.g., see

Bangert-Drowns, Kulik, & Kulik, 1991), both with adults (Campbell & Mayer, 2009; Cranney, Ahn, McKinnon, Morris, & Watts, 2009; McDaniel, Anderson, Derbish, & Morrisette, 2007; though cf. Tse, Balota, & Roediger, 2010, for limitations of testing effects in older adults), and children (Agarwal, Roediger, McDaniel, & McDermott, 2010; Wartenweiler, 2011; though cf. Bouwmeester & Verkoeijen, 2011). Despite the generality of the testing effect, the precise mechanism by which it exerts a mnemonic benefit is unclear. Nevertheless, testing benefits are thought to arise from the process of retrieval, rather than, or in addition to, other byproducts of testing (e.g., see Roediger and Karpicke, 2006a, for discussion of secondary, "mediated" [p. 182] effects of testing).

Both early and contemporary investigations of the testing effect have often compared the effect of testing to a no-test control condition (e.g., Brewer, Marsh, Meeks, & Clark-Foos, 2010; McDermott, 2006; Spitzer, 1913; Wheeler & Roediger, 1992), in which some information to be learned, prior to a final criterion test, is studied then initially tested, while other information is only initially studied. As such, total exposure time to the materials subject to testing becomes confounded with the act of retrieval itself. Early work by Tulving (1967) found that, when equating for acquisition time, engaging in trials of retrieval practice yielded a similar level of learning as do study trials. Similarly, Zacks (1969) manipulated the distribution of tasks (study, covert retrieval, or overt retrieval) for participants to engage in during a set duration learning episode. Results mirrored those of Tulving (1967), suggesting that testing effects were simply an artifact of increased time spent learning test items. Such early results provided evidence of learning occurring during *test* trials; a factor often overlooked at the time (Roediger & Karpicke, 2006a). However, the lack of a benefit of testing over equivalent study failed to support a unique role for retrieval in learning.

Despite early failures to find direct benefits accruing from tests above those of equivalent study, a number of methodological problems (now known to moderate the magnitude- or even emergence- of testings effects) were present (e.g., see Carrier & Pashler, 1992; Kuo & Hirshman, 1996, 1997; Roediger & Karpicke, 2006a). A large body of contemporary research has addressed these problems through the adoption of strict methodological constraints, including the use of a restudy control condition, equating the availability of test and study items in working memory at the time of assessment, and delaying the intervals between subsequent study and test trials to constrain retrieval to long term episodic memory. In such cases, testing advantages reliably emerge, indicating an active effect (i.e., rather than an artifact of exposure time) of retrieval in enhancing retention (e.g., Carpenter & DeLosh, 2006; Carrier & Pashler, 1992; Cull, 2000; Kuo & Hirshman, 1996, 1997; Toppino & Cohen, 2009). However, the driving force behind the memory benefits derived from retrieval has been debated (e.g.,Carpenter, 2009; Kuo & Hirshman, 1996; 1997; Pyc & Rawson, 2010). One approach to explaining the testing effect is to consider the act of retrieval as it relates to dual process theories of memory (e.g., Chan & McDermott, 2007).

Two broad categories of memory models- single and dual process- have emerged to explain what processes can contribute to memory (e.g., see Malmberg, 2008). Single process models have historically had a strong footing within the memory model literature. From earlier global matching models (e.g., SAM, Gillund & Shiffrin, 1984), to modern signal detection based models (e.g., see Dunn, 2004), many authors have argued that a single process, that of memory strength, is sufficient to account for data on recognition memory judgments. Single process theories excel in a variety of areas, including simplicity and parsimony, in that invoking a second process would be

superfluous if data on human memory could be accounted for without that process. Despite this, the prominence of dual process theory has grown in the last 40 years, such that now the majority of cognitive psychologists believe two processes are necessary to explain the existing empirical data pertaining to (at least recognition) memory (Yonelinas, 2002).

Dual process theory posits that recognition memory judgments (see Yonelinas, 2002), cued recall (e.g., Hay & Jacoby, 1996), and free recall (McCabe, Roediger, & Karpicke, 2011) can be informed and influenced via two distinct and independent processes: recollection and familiarity. The former refers to a conscious "remembering" of the study event, in which some form of contextual, qualitative detail can be retrieved along with (in the case of recognition) the memory of experiencing a target item to be recognized. In contrast, familiarity is conceptualized to be a more general feeling of having previously experienced a certain stimulus, but in the absence of any supplementary contextual details from the study event. Similarly, in free recall, familiarity may manifest as item intrusions into one's consciousness without intent (McCabe et al., 2011). Recollection and familiarity are thought to act independently, such that the strength of one is not affected by that of the other. In addition, while not all dual process models prescribe the same characteristics to recollection and familiarity, recollection is generally thought to operate as a threshold retrieval process (i.e., all-or-none), while familiarity is typically characterized as continuous (e.g., Yonelinas, 1999). Dual process theory, because of the empirically testable dissociations predicted between recollection and familiarity given certain types of experimental manipulations, provides an avenue from which the testing effect can be better understood in terms of the specific processes utilized during retrieval.

In order to estimate the contributions of recollection and familiarity during a memory task, a variety of different methodologies can be utilized (e.g., source memory tasks, process dissociation, Jacoby, 1991, etc.). One method particularly amenable to investigating the testing effect is known as the remember-know procedure (Tulving, 1985). During a recognition task, a participant can be asked to provide either a "remember" or a "know" response, reflecting the subjective, phenomenological nature of the act of retrieval for a given target item (see Gardiner, Ramponi, & Richardson-Klavehn, 1998). Participants are instructed to respond "remember" in cases for which they are able to remember the specific episodic event (or at least some component of the event) during which they encoded the target stimulus. As such, remember responses are used as a subjectively judged metric of recollection. Alternatively, participants may respond "know" in the event that the target stimulus evokes a feeling of having been previously experienced, but in the absence of episodic details. Know responses, in this manner, provide an index of familiarity. While the precise contributions of recollection and familiarity underlying remember or know responses have been disputed (see Jacoby, Yonelinas, & Jennings, 1997), it is generally agreed upon that the procedure provides at least approximate process estimates given that certain precautions and interpretation guidelines are followed (e.g., see Chan & McDermott, 2007; Geraci & McCabe, 2006; McCabe & Geraci, 2009).[1]

[1] Remember-know data have been interpreted in the context of both single and dual process models. While the present study interprets remember-know data in a dual process framework, a number of authors have demonstrated the compatibility of such data with single process models (Dunn, 2004; Wixted & Stretch, 2004), where remember and know responses represent relatively higher and lower levels of memory confidence, respectively. Critically, single process models based on memory strength assume that confidence levels (and therefor remember and know responses) vary according to some monotonically increasing function with memory strength.

Testing, as examined in the literature, is by definition an episodic task (Karpicke & Zaromb, 2010). When engaging in a test of memory, one must retrieve information from a previous spatiotemporal episode. That is, information must be recalled from a specific previous time and location, such as an earlier phase of an experimental procedure. As such, testing relies largely on recollection. Testing has been shown to enhance the contribution of recollection (while leaving familiarity constant) to memory judgments when a final criterion recognition test is used (Chan & McDermott, 2007). Early evidence for the testing-recollection relationship comes from Jones and Roediger (1995). Participants learned eight word lists, half of which were subjected to an intervening free recall test. On a later remember-know recognition test over all items, participants recalled the tested items at a slightly higher frequency than the non-tested items. More interestingly, the increase in the old/new recognition hit rate was driven entirely by remember responses. Tested items were given more remember but fewer know responses than non-tested items, indicating a greater contribution of recollection during the final test.

More direct evidence as to the role of recollection in testing comes from Chan and McDermott (2007). Across experiments, whether in the presence or absence of an overall testing effect in recognition hit rates, Chan and McDermott (2007) revealed test enhanced recollection; even when an overall testing benefit is not present in the target hit rate, the act of episodic retrieval alters the available information by which later recognition decisions are made. Converging evidence was provided by the use of a source memory task (Exp. 1a, and Exp. 1b, where testing improved list discrimination), a remember-know recognition task (Exp. 2, where testing increased remember responses) and process dissociation (Exp. 3, modified from Jacoby, 1991, where testing improved performance

on a final exclusion test), with all cases suggesting an increase in the use of recollected details accrued from testing. This finding is especially significant for two primary reasons. First, by virtue of testing, participants were better able to discern which phase of the experiment a previously studied word was presented. In other words, the temporal resolution of the study context was enhanced via testing. Secondly, the phenomenological quality of recognition on the criterion test was modified as a result of testing. When asked to produce a remember-know judgment at recognition, those target items which were tested were more likely to garner a remember response, indicating the presence of conscious recollection of the prior study episode by participants. In sum, both Jones and Roediger (1995), and Chan and McDermott (2007) provide evidence that testing may modulate *how* participants make a recognition judgment, along with *what* details are available from the study episode.

Independent of the testing effect literature, the process of recollection has been shown to be associated with memory of contextual information (Meiser & Sattler, 2007). Spatial judgments as to the location of previously studied target words are enhanced when such words are recognized through self-reported recollection (Perfect, Mayes, Downes, & Van Eijk, 1996). Perceptual details, such as the color of studied materials at encoding, demonstrate a similar benefit driven by recollection (Dudukovik & Knowlton, 2006). Such enhancement of context memory due to recollection seems to occur relatively automatically, regardless of whether a participant is initially told explicitly (Dudukovik & Knowlton, 2006), or not informed (Perfect et al., 1996) as to the nature of the later memory assessment (e.g., a source vs. item memory test).

Despite the positive association between recollection and the retention of context information retention, Brewer et al. (2010), is the only known study to have directly

investigated the potential benefit of testing for context memory (as described in the General Discussion).[2] Yet, recent investigations of the efficacy of testing to benefit related, but non-tested information support, in a broad sense, the proposition that testing may impact memory for more than just target tested information.

The alteration of the processes used during retrieval resulting from testing allows for retrieval practice to enhance memory for more than just target information. Recently, a testing based memory phenomenon aptly titled "retrieval-induced facilitation" ("RIFA," as termed by Chan, McDermott, & Roediger, 2006) has provided some support for the general proposition of non-specific testing benefits. In a typical instantiation of a RIFA experiment (e.g., Chan et al., 2006, Exp. 1), participants are given conceptually related materials to study. Later, a subset of the materials (Set A) undergo a test, additional study, or neither, while the remainder of the materials are not reintroduced (Set B). Results from RIFA studies often find that testing a primary subset of materials (Set A) enhances later retention of materials which were not directly tested (Set B) relative to when the primary materials are alternatively given an equivalent duration of additional study (e.g., Chan, 2009, 2010; Chan et al., 2006; Cranney et al., 2009). In essence, RIFA illustrates the generality of the testing effect to semantically related, but untested information from the study episode.

While RIFA does suggest that the benefit of testing is not entirely target-specific, it only provides direct evidence for enhancement of *semantically related* information

[2] Chan and McDermott (2007) examined the effects of testing on list differentiation (i.e., "what list was this target item presented in?"). List differentiation tasks, which often require temporal context information to successfully complete, can be thought of as tapping into source information (e.g., Johnson et al., 1993). However, such tasks may not tap exclusively into individual, trial-specific context information, but rather may be informed by memory for the general learning episode (i.e., not individual item-specific episodes, e.g., Spencer & Raz, 1995). This difference between list differentiation and other types of source memory tasks prevents the Chan and McDermott (2007) results from being generalized across other source memory dimensions, in which context memory is included (Johnson et al., 1993).

from the study episode, specifically. Furthermore, research on RIFA to date has used only prose (e.g., Chan et al., 2006) and classroom lecture (Cranny et al., 2009) materials, restricting the results from being generalized to all test-enhanced learning. However, some evidence suggests that RIFA may be related to recollection. Chan et al. (2006) found that the magnitude of RIFA varied as a function of time spent producing answers during the intervening test. As reaction times for participants increased, so did the magnitude of RIFA, such that slower participants saw a larger benefit than faster participants in recall of related but non-tested materials relative to non-tested control materials on a final criterion test. A RIFA effect was absent, or even reversed, when participants engaged in relatively fast retrieval, or were discouraged from engaging in a deep, elaborate (i.e., thorough) retrieval strategy during the intervening test. Recollection (in contrast to familiarity) may be a relatively slow process (see Yonelinas, 2002). As the time to complete a recognition decision increases, so does the contribution of recollection to said decision (Yonelinas & Jacoby, 1994). Because RIFA grows alongside retrieval duration, the effect may be mediated in part by enhanced recollection resulting from the intervening test. Such a possibly suggests that the spillover effect of testing may benefit retention of not only information semantically related to the target, but also perceptual or other contextual information which depend on recollected details.

Overview of the Study

Two experiments investigated the effects of testing on memory for information associated with that which has been tested. Experiment 1 assessed the consequences of testing on the retention of perceptual information associated with the tested target that was present at the initial study episode. In addition, Experiment 1 used a remember-know task to replicate past research (Chan & McDermott, 2007; Jones & Roediger, 1995)

which has suggested a relationship between testing and recollection, but while controlling for potential confounds present in the previous studies (e.g., equating exposure time between tested and non-tested material). Experiment 2 used paired associates to examine the influence of testing information on the retention of information conceptually (rather than perceptually) associated with the tested target information. Additionally, Experiment 2 assessed participants' memory for context (the directional consistency of the cue-target pair between acquisition and assessment) as a function of retrieval practice and study.

CHAPTER II

EXPERIMENT 1

Generation, a memory effect similar to testing, has previously been investigated in relation to recollection and context. In a typical generation study, information to be encoded is either "read" by participants (e.g., "NUMBER"), or is self-generated by participants based on an experimenter-provided generation rule (e.g., word completion: "N_MBER"). When memory is later assessed, words encoded via generation are, in many cases, found to be recalled at a greater frequency than those encoded through reading (see Bertsch, Pesta, Wiscott, & McDaniel, 2007, for a review). Using generation as a method of retrieving information during study has been shown to increase the contribution of recollection on subsequent assessment (Yonelinas, 2002). As such, by enhancing recollection, generation may also benefit memory for contextual details of the encoding episode. Consistent with this expectation, certain types of source memory are enhanced by generation (Geghman & Multhaup, 2004). Yet, when considering memory for perceptual details of the study episode, the opposite has been found. Depending on the specific contextual detail being tested, generation either produces no benefit (for contextual features separate from the target item; e.g., background color against which a target is studied), or in some cases may even hinder memory for visual-perceptual and spatial context features (for contextual information integrated into the target; e.g., the target font color) relative to reading (Mulligan, 2004; Mulligan, Lozito, & Rozner, 2006). The generation effect is often thought to be nearly synonymous with testing in terms of

resulting memory behavior, and no theoretical mechanisms have been proposed to distinguish the two effects (Karpicke & Zaromb, 2010). Despite this, memory for contextual-perceptual information following testing is a potential area to expect a divergence between the two effects.

A primary distinction between testing and generation is the type of retrieval that each task requires. Testing taps into memory from a prior study episode while generation bypasses an initial exposure to the target materials and instead requires retrieval from semantic memory at the time of encoding. This distinction underlies the explanation for the absent or negative generation effect for contextual memory advanced by Mulligan (2004) and Mulligan et al. (2006). In observing the effects of generation relative to reading, the type of processing differed for each type of item at encoding. Read items demanded more data-driven, perceptual processing (Jacoby, 1983), while generated items encouraged the use of conceptual processing (see also Blaxton, 1989). As such, Mulligan (2004) argued that perceptual-contextual features of target items were not processed to the same degree for generated, relative to read items. Consequentially, any recollective benefit for contextual memory resulting from generation would be largely limited because of a lack of extensive encoding of perceptual-contextual features during study.

In contrast, testing provides ample opportunity to assess the effect of recollection while not confounding the primary type of processing at encoding (perceptual vs. conceptual) with retrieval (generation). With testing, the processing of target items at initial study is held constant (i.e., all items are "read" in terms of the generation effect literature), and no experimental manipulation occurs until a later, intervening testing phase. Accordingly, any recollective benefit for contextual memory brought about by

testing can be appropriately compared to a comparison condition matched on the type of processing at initial encoding.

The goals of Experiment 1 were twofold. First, I planned to replicate the findings of Chan and McDermott (2007). While Chan and McDermott found enhanced recollection resulting from testing relative to a no-test condition, they did not control for the amount of time tested and non-tested materials were exposed. Due to this, it is possible to account their findings as the result of more exposure for those items which were tested. Perhaps the test condition allowed participants another opportunity to study those items which they were able to successfully recall, which may have been the catalyst of enhancing recollection, rather than the act of retrieval itself. Such an explanation can be addressed by including a restudy control condition, in which non-tested items are presented a second time for additional study (in place of testing), equating the amount of exposure amongst all items in the experiment. Second, in addition to a replication of Chan and McDermott (2007), I sought to examine the relationship between testing target information and memory for perceptual contextual details of the study episode. In particular, Experiment 1 maintains the hypothesis that, by virtue of enhanced recollection, those items which are studied then tested will reveal enhanced memory for perceptual contextual features (in this case, target word font color) present at the initial encoding episode relative to those items that are studied and restudied.

Method

Participants

A total of 64 participants were solicited from a pool of psychology students enrolled at Colorado State University. Participation fulfilled a portion of course

requirements. Data from 6 of the 64 participants were excluded from analysis due to failures in following instructions, yielding data from 58 participants.

Materials

Stimuli were selected by sampling a set of 128 words from Wilson's (1988) database, controlling for the parameters of concreteness (between 200 and 700), and frequency of occurrence (greater than 10 per million). In addition, all items were constrained to under three syllables and between four and nine letters in length. Sixteen lists were generated, each consisting of eight items. Items were randomly assigned to lists, with the constraint that no list contained multiple items with identical first letters. Each item was assigned a color, either red or blue, which specified the font color in the initial study phase of the experiment. Font color was counterbalanced, with each item being presented in each color an equal number of times across participants. Eight of the 16 lists were used as target lists in the experiment, with the other eight lists acting as lures. Target and lure lists were counterbalanced across participants.

Design

The experiment used a mixed-list, within-subjects design. For each of the eight target lists in the experiment, half of the items were subject to an additional study opportunity, while the other half were tested via cued recall (the first letter of the target acting as the cue) during the intervening task. The intervening task served as the independent variable in the experiment. The order of testing and restudying within each list was randomly assigned, with the constraint that no more than two adjacent items were subject to the same type of task. The type of intervening task given to each item was counterbalanced, with each item equally receiving both additional study and testing across participants. In sum, a total of eight experimental instantiations were created,

reflecting the counterbalancing of target versus lure lists, red versus blue font color, and testing versus additional study at the intervening task. The experiment took place exclusively on computers, with each participant tested individually.

Procedure

Participants began by reading two pages of instructions that outlined the procedure of the experiment. Instructions stated that participants would see lists of words that they should commit to memory for a later assessment and that they would be asked to either restudy or be tested on the items in each list (with examples of each condition included). Participants were not made aware of the purpose of the distractor tasks in the experiment. Instructions concluded by allowing participants the opportunity to request clarification from the experimenter on the procedure.

Following the experimental instructions, participants began the experiment. During the initial study phase, every item within each eight-item list was sequentially presented on the screen for 3,000 ms, followed by an inter-stimulus interval of 500 ms. The font color of each item was either red or blue, as specified earlier. Following the presentation of an eight-item list, participants were instructed to mentally compute an arithmetic task for 15 sec then input their answer into the computer.

After the mental arithmetic distractor task, each item from the previously studied list was re-presented sequentially for 5,000 ms in one of two forms during the intervening task phase. For those items presented for additional study, the complete item was shown in black font (e.g., "HOLIDAY"). Alternatively, for those items subject to a test, the first letter of the item was presented in black font (e.g., "H_____"). During the 5,000 ms presentation, participants were asked to type in the word that was presented on the screen or, in the case of tested items, to retrieve and type in the corresponding item from the

previous study phase. Requiring action (typing) of the participant on both test and study trials was used to prevent potential displaced rehearsal effects (Slamecka & Katsaiti, 1987), where participants selectively rehearse test items during the study re-presentation trials. A 500 ms inter-stimulus interval occurred between each item trial during the intervening task. Once the eight-item list was exhausted, participants received a new list and repeated the initial study and intervening task phases of the experiment until all eight target lists were completed.

Following the completion of the initial study and intervening task phases for all eight target lists, participants were given a five min long distractor task, where they were asked to recall as many U.S. states as possible. A final criterion test was given following the long distractor task.

The final test was composed of two steps: target remember-know recognition, and context recognition. Instructions as to the terminology of the recognition responses for the target remember-know recognition task were provided prior to the start of the recognition test. The instructions were based off those used by McCabe and Geraci (2009) (which were derived from Rajaram, 1993), using neutral response terms ("Type A," in place of remember, and "Type B," in place of know) rather than the original remember-know terminology. In addition, source-specific instructions (McCabe & Geraci, 2009) were used, in which Type A instructions specified that the origin of any recollection necessary for a Type A response for a given item should be based on the current experiment only. Using neutral response terms, in conjunction with source-specific instructions, has been shown to provide a more accurate estimate of recollection (McCabe & Geraci, 2009). For the sake of consistency with the larger literature, Type A

and Type B responses will be referred to as remember and know responses, respectively, for the remainder of the the study.

Target and lure items were presented, all in black font, in a random order for the final recognition test, During the target recognition test step, participants made a remember, know, or new judgment for each target and lure item in the experiment. For those items that garnered a remember or know response, a further two-alternative forced choice recognition judgment was solicited as to the color of the target font at initial study (red or blue). After the completion of the recognition test, the experiment halted and participants were debriefed and excused. The experiment lasted approximately 45 min.

Results

The primary analyses were conducted on three variations of the data: unconditional upon performance at the intervening test (UC), or conditional upon either successful (CS, see Runquist, 1983, 1986) or failed (CF) intervening retrieval. Unconditionalized data includes all items in the analyses, regardless of initial retrieval success, while conditionalized data considers only those items in the test condition which were successfully (CS), or unsuccessfully (CF) retrieved during the intervening task. The alpha level was set at $p = .05$. Unless explicitly noted, all statistical tests reported were significant.

Intervening Test Performance

During the intervening task, participants correctly recalled 41% ($SD = .13$) of test items. Performance was comparable with other studies in the literature with similar experimental protocols (e.g., between that seen by Wheeler et al., 1992, and Carpenter and DeLosh, 2006).

Overall Recognition Performance

The data were analyzed using a 2x2 (intervening task: test vs. restudy; memory measure: target vs. context) repeated measures Analysis of Variance. A significant main effect of memory measure, with target memory surpassing context memory, was found in each data set (UC, CS, and CF), $F(1, 57) = 161.95$, 481.44, and 64.89, respectively.[3] There were significant main effects for intervening task in both the CS (test > study) and CF (study > test) data, $F(1, 57) = 37.70$ and 37.39, respectively, along with a non-significant trend towards a overall test item memory advantage in the UC data, $F(1, 57) = 3.42$, $p = .07$. Critically, the intervening task by memory measure interactions were significant in the UC, CS, and CF data, $F(1, 57) = 29.19$, 9.08, 78.28, which were investigated further through planned mean comparisons.

Target Testing Effects

Across all items, the false alarm rate was 11% ($SD = .11$). The target recognition data are shown for each data set in Figure 1. When analyzing the UC data, there was a negative testing effect in the target recognition hit rate, such that more studied items were correctly recognized ($M = .82$, $SE = .02$) than test items ($M = .74$, $SE = .02$), $t(57) = -7.09$. The CF data yielded a similar pattern, with fewer unsuccessfully tested items correctly recognized ($M = .59$, $SE = .03$) than study items, $t(57) = -13.17$. However, the pattern reversed in the CS data, where test performance ($M = .95$, $SE = .01$) surpassed study performance, $t(57) = 9.27$.

[3]The interpretation of the main effect of memory measure should be cautioned as source memory performance was derived only from a subset of all items (those successfully recognized on the target recognition test), while target memory performance was based on all items within a given data set.

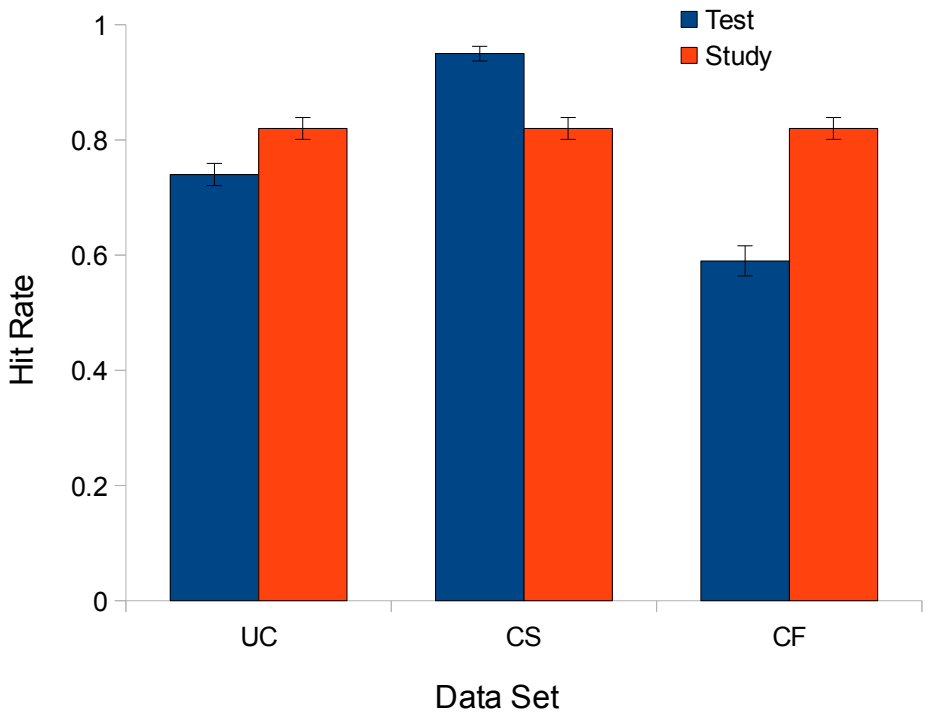

Figure 1. The proportion of correctly recognized target items for each data set as a function of intervening task condition in Experiment 1. The UC data set is unconditionalized, the CS data set conditionalized on successful initial retrieval, and the CF data set conditionalized on unsuccessful initial retrieval.

Remember-Know

The following remember-know results were derived from both raw (i.e., the proportion ofresponses for a given type considering all target items) and conditional (i.e., the proportion of the hit rate) response data (see Chan & McDermott, 2006). Conditional response data help prevent bias stemming from differing target recognition performance across groups. Both types of response data were analyzed across the three variations of data (UC, CF, and CS). Note that descriptive statistics for the raw and conditional remember and know responses within the study condition remained constant across the three data variations. All comparisons across the remember-know response data

remained unchanged after applying a Bonferroni correction of of $a = .05 / n$, where $n = 12$ reflecting the following 12 comparisons. For descriptive statistics see Table 1.

Table 1

Experiment 1 Remember-Know Task Responses

		Remember		Know	
Data Set		Test	Study	Test	Study
UC					
	Raw	0.42 (0.03)	0.41 (0.03)	0.31 (0.02)	0.41 (0.02)
	Conditional	0.57 (0.03)	0.49 (0.03)	0.43 (0.03)	0.51 (0.04)
CS					
	Raw	0.64 (0.03)		0.31 (0.03)	
	Conditional	0.68 (0.04)		0.32 (0.04)	
CF					
	Raw	0.29 (0.03)		0.31 (0.02)	
	Conditional	0.46 (0.03)		0.54 (0.03)	

Note. Values are reported as mean (standard error).

In the UC data, there was no difference in the raw probability of remembering across test and study items, $t(57) = .76$, n.s. However, the conditionalized remember probability was greater for test than study items, $t(57) = 4.02$. Studying lead to a greater probability of know responses in both the raw, $t(57) = -6.64$, and conditionalized, $t(57) = -4.02$, know response data.

The CF data showed a greater raw proportion of remember responses for study items than unsuccessfully retrieved test items, $t(57) = -6.42$. The difference lost significance in the conditional remember data, with remember responses accounting for similar proportions of test and study hits, $t(57) = -1.70$, n.s. The raw proportion of know responses was greater for study than test items, $t(57) = -5.26$, while the conditional know response proportions were not significantly different, $t(57) = 1.70$, n.s.

Considering the CS data, proportionally more tested items were remembered in the raw, $t(57) = 10.37$, and conditional, $t(57) = 8.26$, response data. Conversely, testing led to fewer know responses in the raw, $t(57) = -4.59$, and conditional, $t(57) = -8.26$, response data.

Process estimates of recollection and familiarity were derived from the remember-know response data according to the Independence Remember/Know Procedure (Jacoby et al., 1997). While remember responses are thought to be a relatively pure measure of recollection (one responds remember- and only remember- when recollected details are present), familiarity is thought to contribute to both know and remember responses (i.e., familiarity is often present along with recollection, which is not captured in know responses). As such, using raw response probabilities, recollection was estimated to be equal to the remember responses, while familiarity was estimated as the know responses / (1 – remember responses) (see Jacoby et al., 1997, for more details). The results pertaining to raw remember responses map to recollection estimates, where successful retrieval led to a greater contribution of recollection at final test than did additional study. These results are presented in Figure 2.

However, in contrast with Chan & McDermott (2007), familiarity estimates in the CS data suggested a greater presence of familiarity in successfully tested over studied items ($M_{test} = .79$, $SE_{est} = .04$; $M_{study} = .68$, $SE_{study} = .03$), $t(52) = 2.92$, but the reverse pattern- less familiarity used for tested than study items- in the UC ($M_{test} = .54$, $SE_{est} = .03$; $M_{study} = .69$, $SE_{study} = .03$), $t(57) = -7.07$, and CF ($M_{test} = .44$, $SE_{est} = .03$; $M_{study} = .69$, $SE_{study} = .03$), $t(57) = -10.86$, data. Note that five participants' data were removed from the CS data set due to all targets receiving remember responses, making the Independence

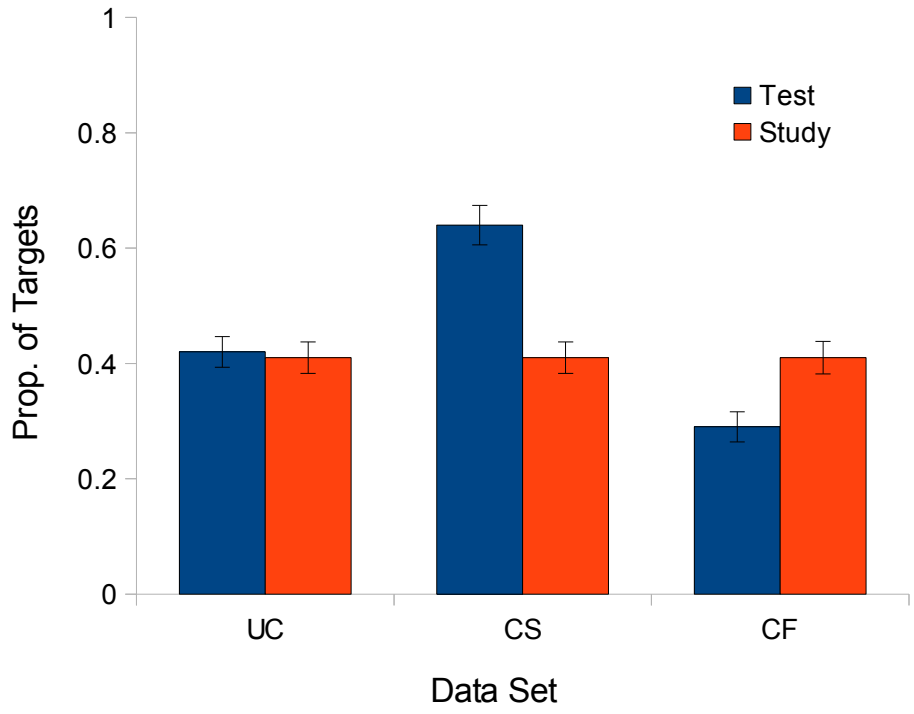

Figure 2. Contributions of remember responses to correctly recalled target items in Experiment 1. Values also double as recollection process estimates (see Jacoby et al., 1997). The UC data set is unconditionalized, the CS data set conditionalized on successful initial retrieval, and the CF data set conditionalized on unsuccessful initial retrieval.

Remember/Know Procedure familiarity estimate incalculable, and thereby likely underestimating the effect magnitude.

Context

Perceptual context memory performance (Figure 3) was computed from the identification-of-origin score (see Johnson, Hashtroudi, & Lindsay, 1993; Mulligan et al., 2006), such that the reported measures represent the proportion of target hits that were attributed to the correct font color. As such, the measure is not biased by overall target recognition performance. Chance performance was 50%. In the UC data, a

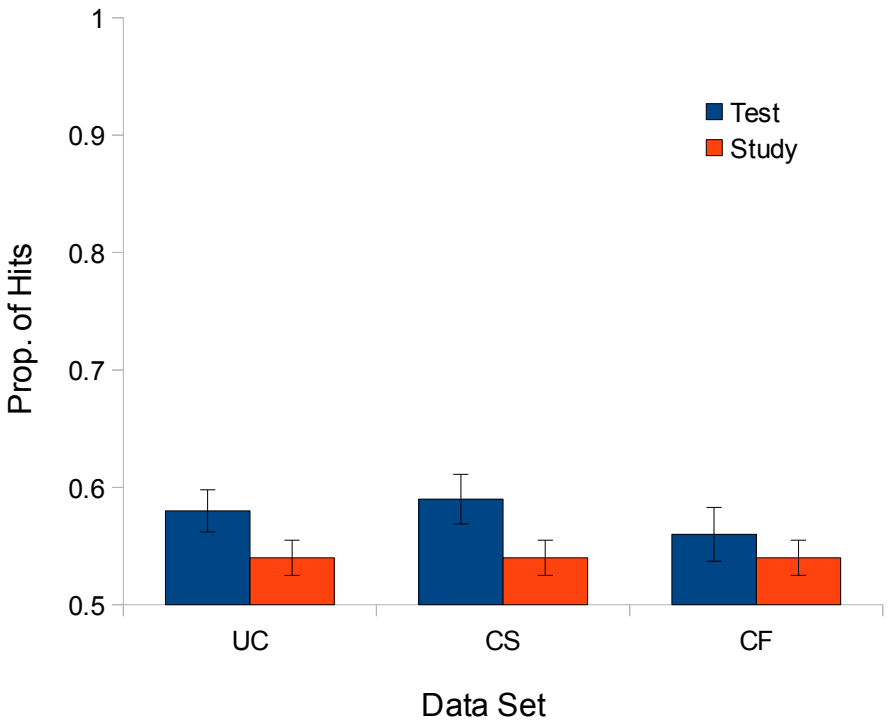

Figure 3. Proportion of target font colors correctly recognized across data sets in Exp. 1 as a function intervening task condition. Chance performance is .5. The UC data set is unconditionalized, the CS data set conditionalized on successful initial retrieval, and the CF data set conditionalized on unsuccessful initial retrieval.

non-significant trend towards a test advantage (M = .58; SE = .02) over study (M = .54, SE = .02) emerged, $t(57) = 1.823$, p = .07. The CF data set failed to yield a difference between study items and test items (M = .56, SE = .02), $t(57)$ = .70, n.s. Yet, when considering successfully retrieved items in the CS data, a significant advantage in context memory performance occurred for test items (M = .59, SE = .02) over study items, $t(57) = 2.02$.

Additionally, individual means from each group and data set (study, UC test, CF test, and CS test) were tested against chance performance (.50). In all cases, performance was significantly above chance, with, respectively, $t(57)$ = 2.71, 4.33, 2.53, and 4.11.

Discussion

Results from Experiment 1 largely aligned with predictions. First, considering the effects of successful retrieval on later target memory, a testing effect emerged, despite the unreliability of detecting testing effects when a final recognition test is employed (see Chan & McDermott, 2007). While the unconditionalized data set found a study advantage in recognition hit rates, the study advantage was driven primarily by unsuccessfully recalled test items reducing the overall test performance mean (see the CF data set).

Performance on the remember-know task replicated the general trends found by Chan and McDermott (2007) and Jones and Roediger (1995). Controlling for overall target recognition performance, test items successfully retrieved at the intervening task saw a greater contribution of remember responses towards target recognition compared with study items. Not surprisingly, unsuccessfully initially retrieved test items evoked the smallest proportion of remember responses during target recognition. In sum, the results suggest that engaging in successful retrieval practice leads to a greater contribution of recollection in later recognition decisions. Although not the focus of the present study, it is worth noting that familiarity estimates were slightly higher for the test items compared with study items, as well.

Central to the primary hypothesis of the study, successfully retrieved test items exhibited greater font color recognition performance than did study items. Cohen's *d* was used as an effect size measure to assess practical differences in context memory performance as a function of intervening task. The formula for *d* derived by Dunlop, Cortina, Vaslow, and Burke (1996) was used to estimate the size of the testing benefit in context memory, where $d = t * (2 (1 - r) / n)^{1/2}$, and *r* is the correlation between pairs

of measures. The Dunlop et al. (1996) *d* measure provides a more accurate estimate of effect size for correlated (within-subject) data sets, and as such, was used in preference to the traditional *d* as defined by Cohen (1988). c Despite not reaching significance, the magnitude of the trend towards a test advantage in the UC data ($d = .30$) nearly matched that of the significant testing effect in the CS data ($d = .34$). This pattern of results was influenced by unsuccessfully retrieved test items performing, unlike in target memory, just as well as (or slightly better than; $d = .12$) study items on the font color recognition test.

Interestingly, engaging in an intervening test seemed to dissociate memory performance on the target and context recognition tests. Considering UC target recognition performance, an overall negative testing effect emerged, driven by the unsuccessfully retrieved items. Yet, there was a de-facto positive testing effect (see above effect sizes) in UC context memory performance. In contrast to the target test, unsuccessfully retrieved items were recognized equally as well as study items in the context memory task, thus demonstrating a crossover interaction between performance on the two memory measures by item type (see overall recognition performance analyses above).

.CHAPTER III

EXPERIMENT 2

While Experiment 1 examined the effects of testing on memory for perceptual context information, Experiment 2 focused on conceptual, associative context. In particular, Experiment 2 investigated how testing a piece of information influences memory for other non-tested, but conceptually associated information. Cue-target word pairs were used to establish associations to be tested.

A large body of literature has garnered support for the idea that cue-target pairs are represented holistically in memory (e.g., see Kahana, 2002). A popular instantiation of this idea has been referred to as the Principle of Associative Symmetry (Asch & Ebenholtz, 1962). According to the Principle of Associative Symmetry, a pair of items, such as a cue-target word pair, becomes represented in memory as a unitized, holistic piece (i.e., each word pair member is bound together in memory as a single episodic event). As such, recall of either item within a pair is not thought to be dependent on the specific directionality of the association (i.e.., given an A-B pair, A - _ vs. B - _).

Of particular relevance to the present study, recent evidence has shown that associations can be holistically sensitive to the individual item characteristics of their constituent members (Madan, Glaholt, & Caplan, 2009). For example, among sets of matched and mismatched high and low imageability cue-target pairs (high-high, high-low, low-high, low-low), recall performance for both the target (given the cue) or the cue (given the target) are dependent on the item characteristics of both items in the pair,

rather than just the item being tested (Madan et al., 2009). Retrieval practice is thought to produce mnemonic benefits in part due to an enhancement of item-specific processing (e.g., see Karpicke & Zaromb, 2010). As such, it may be expected that engaging in a test of memory for one item within a cue-target pair may not only yield a testing effect for that item itself, but also may benefit memory for the associated item.

Preliminary evidence from Carpenter, Pashler, and Vul (2006) supports this conclusion. Carpenter et al. (2006) had participants study A-B word pairs. Later, some of the word pairs were presented for additional study, while others were subject to a A - ? cued recall test. On a final criterion test, participants were asked to recall words in the same direction as the initial testing opportunity (A - ?), or in the opposite direction (? - B). Regardless of the order of the final test, a testing effect was found in all cases, such that initial A - ? testing yielded a memory advantage for both later A - ? and ? - B recall. The result appeared robust across both final cued and free recall tests. While in support of the Associative Symmetry Principle, Carpenter et al. (2006) provided feedback during the intervening task after each trial. Under at least certain experimental conditions and retention measures, the inclusion of feedback yields benefits above those seen with pure testing in isolation (e.g., Butler, Karpicke, & Roediger, 2007; Kang, McDermott, & Roediger, 2007; also see Cranney et al., 2009). The inclusion of feedback is confounded with testing, and as such, it is unclear what precisely contributed to the purported testing benefit.

Similar research by Sommer, Schoell, and Buchel (2008) examined the effects of testing on memory for visual-spatial associations. Participants learned image-location associations for 16 items in a 4 by 4 grid of locations. During an initial test phase, for half of the stimuli, participants were either given an image and asked to select the correct

location it had appeared in earlier, or given a location and asked to identify the corresponding image from the initial study phase. Participants were allowed to select more than a single image or location on each trial if they were unsure. Later, a second test was given, in which all image-location associations were tested, either in the same or opposite direction of earlier testing. Of particular interest, when only considering those responses in which the participant provided a single response (akin to being able to write only one word during a verbal memory test), a testing effect was found. The testing effect did not differ across the direction of association, and was the same for both congruent (e.g., test 1: image-?, test 2: image-?) and incongruent (e.g., test 1: image-?, test 2: ?-location) tests. Despite the strong evidence provided by Sommer et al. (2008) of information being strengthened holistically through retrieval practice, a restudy control condition was not employed. As such, memory for associations could be explained as potentially arising from more exposure to image-location pairs for some items during the first retrieval opportunity. In addition, the nature of the materials used (images and locations) differ from the the current study, and as such, the results should not be automatically generalized to verbal associative information.

While Carpenter et al. (2006) and Sommer et al. (2008) both provide evidence in support of test-enhanced associative memory, the mechanisms giving rise to this result is unclear. Rather than, or in addition to, the representation of associated pairs of stimuli being treated holistically in memory, test-aided recollection may contribute to memory between associated items. That is, if testing one item in a pair promotes recollection, gains in memory for the whole pair of stimuli may be in part the result of a reinstantiation of the study context in which both items were present. Experiment 2 investigated this possibility.

Participants studied paired associates (in the form A - B), after which point some pairs were presented for additional study, while others were given a forward directional cued recall test (A - ?). On a subsequent final test, all pairs were tested, with half in the forward direction (A - ?) and half backwards (? - B). In addition, after each cued recall trial, participants made a source memory judgment, identifying the directional consistency of each word pair between initial study / intervening task and final test (same or reverse). It was expected that the results would replicate Carpenter et al. (2006), but without the inclusion of feedback during initial testing. The source memory task provided a measure of retention for association directionality: a component of the context experienced during study by the participants. It was hypothesized that a generalized test-based enhancement in memory for all items regardless of directional congruity across the initial and final tests would be found. Because recollection has been pinned as the critical process in associative memory (Hockley & Consoli, 1999; Yonelinas, 1997; though cf. Quamme, Yonelinas, & Norman, 2007; Yonelinas, Kroll, Dobbins, & Soltani, 1999), a test derived advantage in the source memory task was also expected via enhanced recollection.

Method

Participants

Seventy-three participants were solicited from the Colorado State University Psychology Research Pool. Participation served as a partial fulfillment of course requirements.

Materials

Wilson's (1988) database was used to randomly sample 224 words, controlling for concreteness (between 200 and 700), and frequency of occurrence (greater than 10 per

million). All words selected were also constrained to between four to nine letters in length, and one to two syllables. The 224 words were randomly assigned into groups of two, creating 112 pairs, allowing for 14 lists consisting of eight word pairs each (randomly sampled from the 112 pairs) to be established. Within each of the 14 lists, four word pairs were randomly assigned to the study condition, and the other four to the test condition. Each word pair was represented as an A - B pair, with the "A" and "B" words in each pair reversed in an experimental counterbalance.

Design

The experiment utilized a mixed-list, within-subjects design. The primary experimental manipulation occurred during the intervening task, in which test pairs were subjected to a cued recall test (in the direction A - ?), while study pairs were re-presented for additional study. Test and study pairs were counterbalanced across participants. A total of four experimental instantiations were used, reflecting the test-study counterbalance and the A - B word pair order counterbalance. Participants were tested individually on computers.

Procedure

The experiment began by having participants read two pages of instructions outlining the procedure of the experiment. After reading the instructions, participants entered an initial study phase for the first list of word pairs. Each word pair within the list was presented sequentially on the screen in the form A - B for 3,000 ms, (e.g., "APPLE - GLASS") followed by a 500 ms inter-stimulus interval. Following the presentation of the eight word pairs, participants moved on to the first distractor task, in which they were asked to perform a mental math task for 15 sec and record their answer.

Following the first distractor task, participants entered the intervening task phase of the experiment. Each of the eight word pairs within the current list were re-presented either in a complete (study group) or partial (test group) form. For those pairs that were presented again completely (e.g., "APPLE - GLASS"), participants were asked to restudy the pair and type in the B word as seen on the screen (GLASS). For those word pairs which were represented in incomplete form, the A word appeared on the screen followed by a blank ("APPLE - _____"), during which time participants attempted to retrieve the B word (GLASS) of the pair and type it into the computer. Each pair appeared on screen for 5,000 ms, followed by a 500 ms inter-stimulus interval.

After the intervening task phase of the experiment, participants received a new list and repeated the above procedure until the exhaustion of the 14 experimental lists. Following the completion of the intervening task for the 14th list, participants were given a long distractor task. For five min, participants were asked to recall as many U.S. states as possible. After the long distractor task completed, participants began the final test phase of the experiment.

During the final test phase, participants were given both a cued recall test and a source memory test in succession for each word pair. Test ordering was random across all 112 word pairs. For the cued recall test, a word from each pair appeared on the screen followed by a blank. A random half of the cues were the A words (A - _), while the other half were the B words (B - _). Participants were asked to retrieve and type in the corresponding member of each pair given the cue if they were able to recall it. Following the cued recall test for a word pair, participants were given a two-alternative forced choice test as to the directionality of the word pair just presented. For each word pair, the cue (as presented during the immediately previous cued recall test, as either A or B)

appeared on the screen along with the correct target (regardless of whether the participant successfully recalled the target item). Participants were then asked whether the cue-target pair, as presented, was in the same direction (A - B) as studied earlier in the experiment, or the reverse direction (B - A). The duration of the cued recall and source memory tests was participant-paced. Following the completion of the final test phase, participants were debriefed and excused. The experiment lasted approximately 50 min.

Results

The alpha level was $p = .05$. Similar to Experiment 1, analyses were conducted for the target and source memory measures on data unconditionalized (UC), conditionalized on successful intervening test retrieval (CS), and conditionalized on failed intervening test retrieval (CF). Two additional data sets are introduced for the source memory task analysis: CCS and CCF. The CCS data set considers source memory performance only for those items which were successfully retrieved on the *final* target cued recall test. Conversely, the CCF data set takes into account only those items failed to be retrieved on the final target cued recall test. Note that, in both the CCS and CCF, the subset of study items differed from those of the UC, CS, and CF data sets, as both test and study items could be conditionalized upon final test retrieval (while only the former can be conditionalized on initial test retrieval). All statistical tests are significant unless explicitly noted otherwise.

Intervening Test Performance

Participants successfully recalled 47% of the test targets correctly at the intervening test. Split into conditions reflecting the later directional congruity between

initial and final tests, congruent (M = 0.48, SE = .02) and incongruent (M = 0.45, SE = .03) items were recalled at similar frequencies, $t(72) = 1.81$, n.s., suggesting comparable mean item difficulty across congruency conditions.

Cued Recall Target Testing Effects

A 2x2 (intervening task: test, study; directional congruity: same, reverse) repeated measures Analysis of Variance assessed the testing effect across directionalities (Figures 4-6). See Table 2 for descriptive statistics. In the UC data, there were significant main

Table 2

Experiment 2 Final Target Cued Recall Test Performance

	Test Items		Study Items	
Data Set	Same	Reverse	Same	Reverse
UC	0.27 (0.02)	0.22 (0.02)	0.17 (0.02)	0.16 (0.02)
CS	0.49 (0.03)	0.37 (0.02)		
CF	0.05 (0.01)	0.08 (0.01)		

Note. Values are reported as mean (standard error).

effects of both intervening task, $F(1, 72) = 58.99$, $\eta_p^2 = .45$, and directional congruity, $F(1, 72) = 14.16$, $\eta_p^2 = .16$, with overall recall for tested items surpassing that of studied items, and same direction performance exceeding reverse direction performance. Additionally, the intervening task by directional congruity interaction was significant, $F(1, 72) = 9.57$, $\eta_p^2 = .12$. Follow-up comparisons showed superior test performance in both the same, $t(72) = 8.21$, and reverse, $t(72) = 5.23$, directionality conditions.

The CS data set results mirrored those of the UC data set, with significant main effects of both intervening task, $F(1, 72) = 323.13$, $\eta_p^2 = .82$, and directional congruity, $F(1, 72) = 19.78$, $\eta_p^2 = .22$, along with a significant interaction, $F(1, 72) = 20.02$, $\eta_p^2 = .22$. As in the UC data, the testing effect was larger in the same compared to reverse

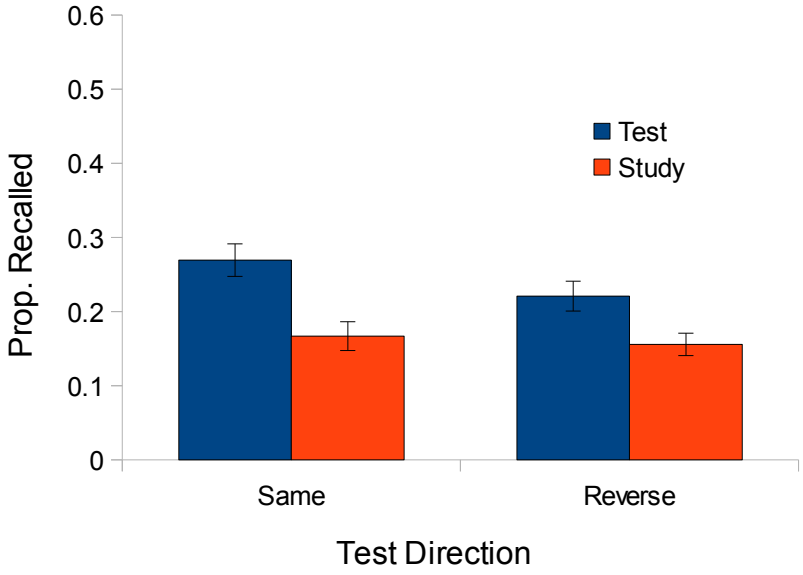

Figure 4. Performance on Experiment 2 final cued recall task in the UC (unconditionalized) data set. Performance shown as a function of intervening task condition and directional congruity of word pair between initial study/intervening task and final test.

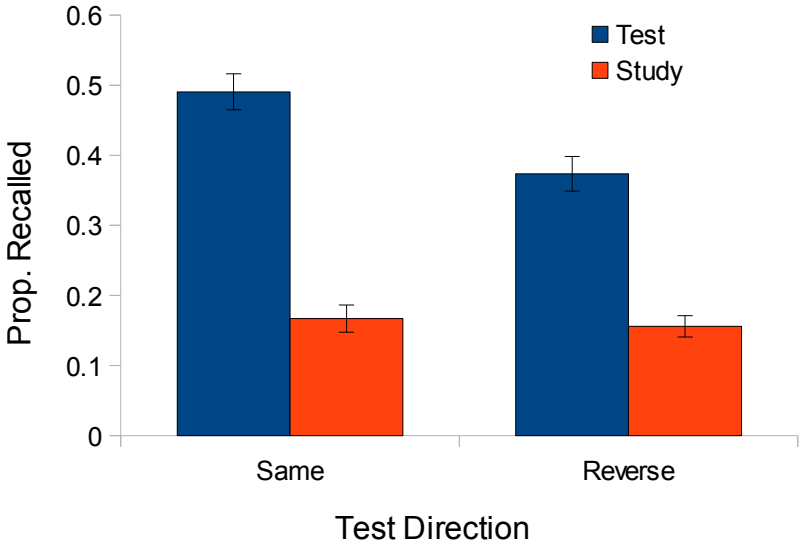

Figure 5. Performance on Experiment 2 final cued recall task in the CS (conditionalized on successful initial retrieval) data set. Performance shown as a function of intervening task condition and directional congruity of word pair between initial study/intervening task and final test.

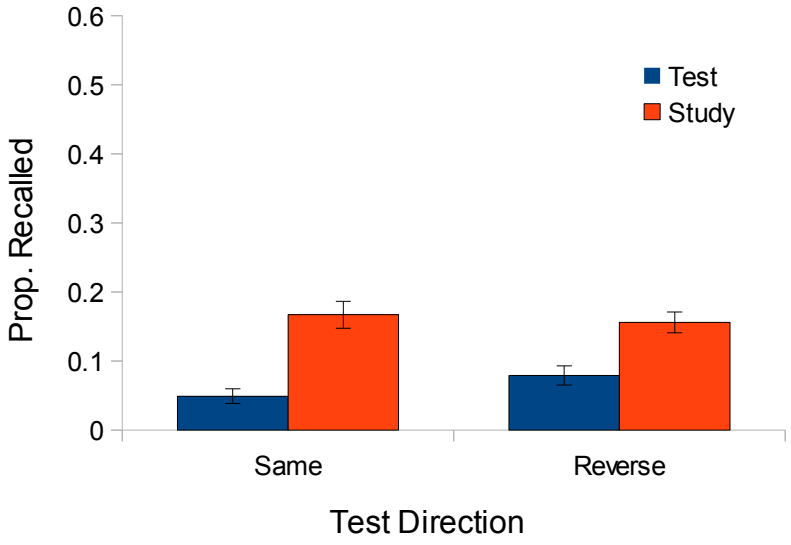

Figure 6. Performance on Experiment 2 final cued recall task in the CF (conditionalized on unsuccessful initial retrieval) data set. Performance shown as a function of intervening task condition and directional congruity of word pair between initial study/intervening task and final test.

direction condition, though significant in both cases, $t(57) = 17.86$, and 10.81, respectively.

The CF data yielded a significant main effect of intervening task, $F(1, 72) = 90.95$, $\eta_p^2 = .56$, with study items outperforming test items. The main effect of directionality was not significant, $F(1, 72) = 1.21$, n.s., though the interaction between the two factors was, $F(1, 72) = 5.00$, $\eta_p^2 = .07$. Unlike the CS and UC data, unsuccessfully retrieved (CF) items performed poorer than study on the final cued recall test for the same, $t(72) = -8.63$, and reverse, $t(72) = -5.58$, though the study advantage was mitigated in the latter.

Source Memory Directionality Test

Source memory test performance (Figure 7) was assessed for the UC, CS, and CF data, along with the two additional data sets: items successfully (CCS), or unsuccessfully (CCF) retrieved on the final cued recall test.

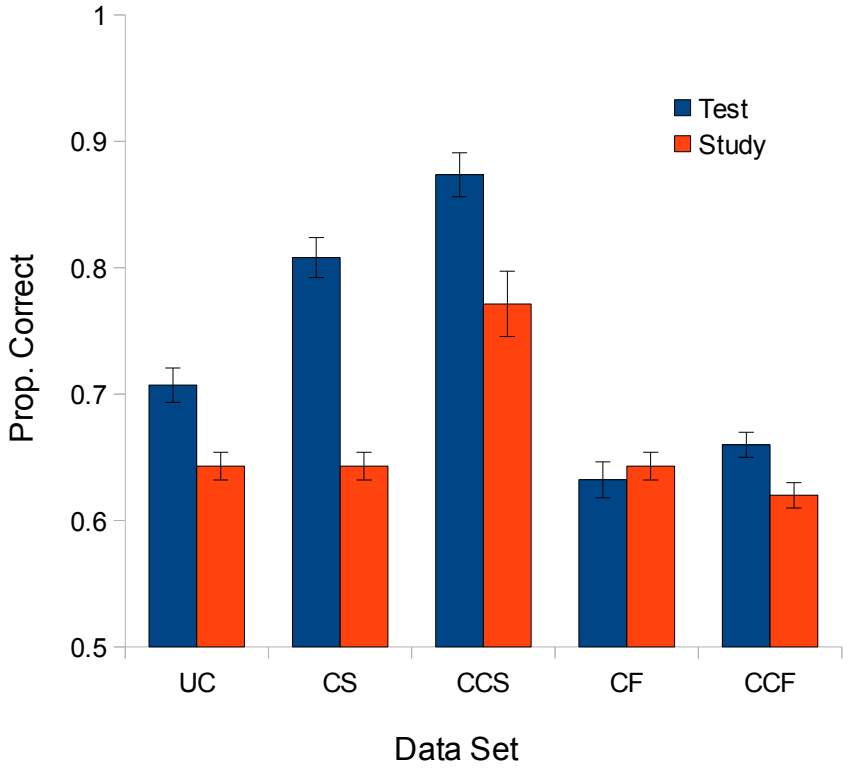

Figure 7. Performance on the directionality test in Experiment 2. Proportion of correct judgments of word pair directionality shown as a function of intervening task condition and data set. The UC data set is unconditionalized, the CS data set conditionalized on successful initial retrieval, and the CF data set conditionalized on unsuccessful initial retrieval. The CCS data set is is conditionalized on successful final cued recall test retrieval, and the CCF data set is conditionalized on unsuccessful final cued recall test retrieval.

The UC, CS, CCS, and CCF data sets all produced the same pattern of results, with test items (M_{UC} = .71, SE_{UC} = .01; M_{CS} = .81, SE_{CS} = .02; M_{CCS} = .87, SE_{CCS} = .02; M_{CCF} = .66, SE_{CCF} = .01) outperforming study items ($M_{UC\text{-}CS\text{-}CF}$ = .64, $SE_{UC\text{-}CS\text{-}CF}$ = .01; M_{CCS} = .77, SE_{CCS} = .03; M_{CCF} = .62, SE_{CCF} = .01) in each case, $t(72)$ = 5.62, 12.18, 3.57, and 3.27, respectively. In the CF data, no differences emerged between study and test conditions (M_{CF} = .63, SE_{CF} = .01), $t(72)$ = -0.77, n.s.

Discussion

Similar to Experiment 1, Experiment 2 demonstrated a generalized benefit of successful retrieval not only for target recall, but also for information associated with that which was tested. Testing improved performance on target tests of either directionality. The same pattern of results emerged in both the UC and CS data sets, though with a larger magnitude testing effect in the latter. Thus, superior test item recall was driven predominantly by items successfully retrieved at the intervening test. Furthermore, the data provide a replication of the bidirectional testing benefits shown by Carpenter et al. (2006), though without the potentially confounding influence of post-initial test feedback.

Across the UC and CS data, there was a larger testing effect in the case of same direction items (i.e., the same item was the target at intervening and final tests). Furthermore, the effect was driven almost exclusively by a drop in test item, rather than study item performance in the reverse direction condition. This finding suggests that, while test induced advantages in memory carry over to associated but non-tested items, the benefit is slightly reduced.

Even though there was an overall study advantage in target cued recall compared with unsuccessfully retrieved test items, the CF test items performed slightly better in the case of reverse tests (compared with CF test items in the same direction). Considering

that target items on the final test in the reverse condition had received additional exposure relative to the unsuccessfully retrieved intervening test target, the small boost in reverse direction test performance was likely due to total exposure time rather than any beneficial effects of failed retrieval.

The target cued recall test results, on the whole, suggest that test-enhanced recollection contributed to a testing effect in the reverse direction condition. On the other hand, test-enhanced recollection, in addition to any test-induced target-specific advantages, were able to contribute additively in the same direction condition, thereby resulting in a larger magnitude testing effect. In the case of study items, both the cue and target received equal exposure, suggesting that less active processing (i.e., typing the target word as presented during additional study at the intervening task, rather than retrieving the target) does not induce any target-specific advantages and has little mnemonic value.

Extending the findings from Experiment 1, source memory was strengthened through testing. Three key observations emerged from those data. First, the UC data show a testing effect in source memory recognition performance, despite mediocre initial test retrieval success. Second, as in Experiment 1, evidence of a dissociation between target and source memory emerged, where unsuccessfully retrieved (CF) test items yielded a negative testing effect in target recall, but an absence of an effect in source recognition. This null effect occurred despite the re-presentation of the 'A' word in each pair during the intervening test. Theoretically, a second presentation of the 'A' pair member could serve to improve performance on the source task (as the position of only one word pair member is necessary to correctly make the source judgment). Thus, the

lack of a CF source testing effect may provide evidence of the validity of the source test as tapping into source, rather than item memory.

A third key observation came from the CCS and CCF data. Both the CCS and CCF data sets showed a positive testing effect in source memory. By only considering those items which were successfully (or unsuccessfully) recalled on the final test (applied to both item types), differential item memory effects between study and test items were mitigated. As such, the CCS and CCF source memory analyses lend additional support to the validity of the directionality test as a source memory measure, rather than simply reflecting differing item memory for either or both constituent word pair members.

CHAPTER IV

GENERAL DISCUSSION

The broad goal of the present study was to investigate how memory is modified by testing, not only for target material that is itself tested, but also for associated and contextual information. To this end, both experiments found support for testing benefits that extend beyond, or spill over from, that information which itself was tested.

Evidence from the data suggest that test-enhanced recollection may have been, in part, a driving force behind the observed target and context memory testing effects. Recollection (R) estimates (raw remember responses) were related to context memory task performance effect size in Experiment 1, such that in the CS data (test item R_{est} = 0.64), a significant testing effect was present, with the magnitude of the effect decreasing in the UC data set (R_{est} = 0.42) and even more so in the CF data set (test item R_{est} = 0.29). Similarly, when the process estimates for each given data set type in Experiment 1 are mapped to the analogous data sets (UC, CS, CF) in Experiment 2, the same pattern emerges in the directionality source test.

However, test induced recollection was not a pure predictor of context recognition performance. CF test items received significantly lower estimates of recollection than did study items, yet this did not yield poorer CF test item context recognition performance in either experiment. While this finding may be argued as reflecting a floor effect obscuring any differences in context memory, performance was significantly above chance for both test and study items across all data sets in both experiments. Rather, information beyond only that embedded in recollected details (as captured through the remember-know task) may have contributed to context memory, or the intervening task

re-presentation of study items may have interfered with later context memory recognition, thereby masking a study advantage in the CF data set.

The testing effects in source tasks found across both experiments contrast with the only other study (to my knowledge) to have explicitly investigated the topic. Brewer et al. (2010) had participants study two lists, in which words were concurrently presented visually on a computer screen and acoustically in a male or female voice. The test condition included an immediate free recall test for the targets (Exp. 1), the targets along with each associated gender (Exp. 2), or the targets plus the corresponding inter-list position (beginning, middle, or end; Exp. 3) after each list presentation. A no-test control condition included a math distractor task with equal duration to the test condition initial free recall opportunity (Exp. 1 and 2 only). A final recognition test, either over list discrimination (i.e., temporal context: "which list was each target presented in?") or gender discrimination (i.e., perceptual auditory context: "what gender was the speaker of the target acoustic presentation?") was given immediately following the second list free recall / math distractor task (depending on condition). The results indicated that intervening testing improved final recognition performance in cases where the final source task dimension was overtly and intentionally retrieved at the intervening test. Initial free recall testing of the targets alone (Exp. 1) led to better list discrimination, but not when gender information was also solicited during the initial free recall test (Exp. 2). However, the gender and target test group saw superior gender- but not list-discrimination final source test performance. Similarly, inter-list position recall practice (Exp. 3) led to poorer performance on list discrimination than did plain free recall. In essence, Brewer et al. (2010) suggested that testing effects in source memory are critically dependent on retrieval practice tapping into the relevant source dimension, and

are thus not generalizable across different source tasks. Their findings can be framed neatly within a transfer-appropriate processing framework (e.g., see deWinstanley, Bjork, & Bjork, 1996; Morris, Bransford, & Franks, 1977).

The results of the present study are at odds with those of Brewer et al. (2010). While it may be argued that the intervening test of Experiment 2 in the present study tapped into the source characteristic (directionality) which was later tested, participants were not instructed to attend to or specifically retrieve this information. Experiment 1 did not require or request any information from the participants concerning context information (font color) during initial testing. Regardless, a testing effect in context memory was found in both cases. The disparate results between the two studies may, perhaps, be explained by fundamental differences in the methodologies and designs used. Brewer et al. presented to-be-learned information in both visual and auditory modalities. Yet, only one modality- auditory- carried the target source information. Visually-based encoding can enhance the availability of recollective details (i.e., distinctive features) of target information to a greater degree than auditory-based encoding (Pierce & Gallo, 2011). Furthermore, in the case of Brewer et al.'s initial free recall test, retrieval was not constrained or cued to either modality in particular. Thus, when not explicitly instructed to retrieve source information during retrieval practice (i.e., the free recall only group), participants had no particular strategic reason (i.e., agenda, Johnson, 1992; Johnson et al., 1993) to recollect contextual information pertaining to the auditory, rather than visual initial target presentation, specifically (the latter of which matched the initial test output modality, in which participants typed into a computer and viewed, rather than verbalized their retrieved targets).

This account does lend support to Brewer et al.'s conclusion of specific, rather than general recollection-derived test benefits, but only with the assumption that auditory source memory was successfully encoded to any meaningful degree at initial study. In other words, one must assume that Brewer et al.'s participants attended to or otherwise encoded source information from the auditory target presentation in particular, while in the current study, relevant source information was embedded within the single modality of presentation. Even so, modality match effects occur in source recognition memory (Mulligan, Besken, & Peterson, 2010; Mulligan & Osborn, 2009), where a mismatch between study and test modality can reduce a source test hit rate. Recognition hit rates from Brewer et al. may have been further reduced due to the mismatch of target and source information being studied auditorily then cued visually for the final source test. Despite differences in initial item presentation and potential modality (mis)match effects between the two studies, such an inconsistency may be moot given the nature of Brewer et al.'s retrieval practice phase.

Much research points to the necessity of initial retrieval difficulty in producing a later testing effect (and the magnitude of the effect), whether through increasing lag prior to the initial test onset (e.g., Karpicke & Roediger, 2007; Modigliani, 1976; Pyc & Rawson, 2009; Whitten & Bjork, 1977), providing fewer retrieval cues (e.g., Carpenter & DeLosh, 2006), or other means (e.g., Gardiner, Craik, & Bleasdale, 1973; Pyc & Rawson, 2009; also see Roediger & Butler, 2011). Similarly, many have argued that testing advantages only emerge, or become larger, after an increasingly long final retention interval (e.g., Carrier & Pashler, 1992; Roediger & Karpicke, 2006b; Runquist, 1983, 1986; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003).[4] Testing effects in

[4] Rowland and DeLosh (unpublished) demonstrated testing effects after short (one min) retention intervals

non-target information (RIFA in particular) are critically dependent on the use of a slow and broad rather than fast and narrow retrieval attempt at initial test (Chan et al., 2006). Low-effort retrieval of readily available information during an immediate test promotes use of the latter. Brewer et al. used a design with immediate initial and final tests, where initial testing may not have been sufficiently difficult to produce testing benefits in non-target memory. Furthermore, the lack of any delay prior to retrieval practice likely precluded some retrieval attempts from accessing long term memory at all (i.e., short term memory recency effects may have occurred). Indeed, across all statistical comparisons, Brewer et al. found no difference in their results when considering either the successfully (i.e., CS), or unsuccessfully (i.e., CF) initially retrieved test items in their analyses. In sum, Brewer et al.'s design may have created conditions adverse to the emergence of testing effects in source memory.

While the above noted features of Brewer et al.'s methodology may help explain the null testing effects in source memory performance, they do not address the significant testing effects that occurred when the source dimension was initially tested along with the targets. However, a no-test condition was used for control, rather than a restudy condition. Initial (low effort) testing may have been analogous to a second study opportunity for retrieved items when they were recalled along with source information (in addition to promoting the binding of source information to the visual, rather than acoustic, target representations as they were output on the initial test), thereby promoting

for those items successfully recalled during the initial test. However, the inclusion of even a short (e.g., one min) interval prevents recency effects from short term memory from clouding the final test results. In addition, the effect was only found for specifically those items successfully retrieved initially (CS items), while Brewer et al. (2010) found no distinctions between CS and CF items across all experimental comparisons.

encoding variability (McDaniel & Masson, 1985) of the test condition targets and associated sources.

Despite methodological differences, the present study and Brewer et al., taken together, demonstrate the critical moderating role of experimental design factors in the testing effect. Boundary conditions seem to apply, with testing effects in source memory appearing sensitive not only to the act of retrieval itself, but also the processing engaged during learning (i.e., *how* information is encoded at the study episode, combined with *what* participants attempt to consciously retrieve about the study context, combined with the completeness (Glover, 1989) of the retrieval practice event itself). Taken together, the disparate results from the present study and Brewer et al. suggest a promising area for further investigation to identify the limits and factors critical to test-influenced source memory.

Theoretical Implications

An unanticipated finding emerged across the two experiments, where source memory performance for test items never fell below that of study items across every data set. Of specific interest were the results from the Exp. 2 source memory test. Compared to initially unsuccessfully retrieved (CF) items, study items saw no benefit in source memory performance (i.e., a second presentation of items seemed to provide no later advantage in source recognition memory). Yet, comparing study to test items while constraining the analysis only to those items not retrieved on the final cued recall test (CCF), initial testing led to a source memory advantage.

A set of preliminary conclusions can be drawn from this pair of analyses. From the CF data, failure of initial test recall led to large item memory (i.e., target cued recall performance) deficits, though once the target was re-presented to the participant (recall

that the cue-target pair was re-presented to participants for the source recognition test regardless of final target test retrieval success), source memory performance for CF test items returned to that of the study item condition. Cook, Marsh, and Hicks (2006, Exp. 1) investigated the effects of successful versus unsuccessful target cued recall with paired associates on a source memory task either in the presence or absence of the target. After initial study of auditorily presented paired associates, participants took a cued recall test followed by a source memory test (gender of the speaker for the item during initial study) in the absence of any additional information (i.e., feedback) provided by the experiment. For unsuccessfully retrieved target items, participants were then immediately instructed to select the target out of 4 possibilities (3 previously unstudied lures plus the target) and were given an additional source recognition opportunity. Source memory performance for unsuccessfully retrieved items was at chance during the initial source judgment (M = 0.50), but after the re-presentation of the target (amongst lures), performance jumped slightly above chance (M = 0.56). Similarly, after strengthening the binding between cue, target, and source by using three, rather than one initial study trial (with the source information present each time), the effect was strengthened (M = 0.57 before target presentation, M = 0.66 after, Cook et al., 2006, Exp. 2).[5]

Considering these findings, in the present study neither the initially unsuccessfully retrieved (CF) test items nor study items likely received any benefit in source memory from the intervening task experiment phase. But, while item memory was slightly boosted by the intervening re-presentation for only study items relative to CF

[5]Cook et al. (2006) include an additional manipulation of incidental versus intentional encoding of the source information. I only report their incidental encoding condition, given that this condition was most comparable to the present study, since there was no mention of the source directionality test in Experiment 2. Also of note, Cook et al. constrain their analysis of source memory performance on the second source test for unsuccessfully retrieved items (i.e., the test following the re-presentation of the target amongst lures) only to those items which were correctly recognized amongst the 3 lure distractors.

test items (as seen in final target cued recall results), a similar level of *binding* may have occurred during initial study between context and target item information for both CF test and study items, with no reinforcement of the binding occurring during the intervening task in either case. This account would predict, as the data show, poorer item but equal source memory performance for CF test versus study items, respectively, specifically when the target is revealed during the source judgment. Put another way, the target in each word pair may serve as a cue to the associated source information embedded within the original study episode.

The analysis from items unsuccessfully retrieved at the final test (CCF) lend additional support to this account. Inherently, CCF test and study items are "equated" on item memory (i.e., target retrievability). Yet, the source memory advantage for CCF test over study items can be thought to reflect an intervening-test induced advantage in binding that persists despite one's failure to retrieve target information at final test. On the surface, the CF and CCF results seem in conflict. They need not be, though, as the CCF items are not specific to intervening test recall success or failure, and as such, consist of both successfully and unsuccessfully initially retrieved test items. Therefore, the CCF data set, like the CCS data set, suggests that successful retrieval practice may boost both item memory and the binding of targets to contexts, though if the dynamics of Cook et al. (2006) apply, test-induced source memory benefits should only be uncovered when one has access to target item information during the source judgment. Extending from this, there is scant support for any positive or negative influence of unsuccessful retrieval practice on source memory relative to a similar amount of study (cf. Kornell,

Hayes, & Bjork, 2009, for demonstrations of positive effects of unsuccessful retrieval on target memory).[6]

The design of Experiment 1 precludes any perceptual context memory data to be analyzed conditional upon final test performance (e.g., analyses on CCS or CCF data sets), as participants only made a font color judgment for those items recognized as old. In this way, the context memory UC, CS, and CF data sets can all be thought of as being additionally conditionalized on successful final test recognition. The CF data set in particular suggests that additional study of a target provides no advantage over unsuccessful initial retrieval in later context memory performance. Conversely, successful retrieval (CS data set), boosts context memory relative to an equivalent duration of study after excluding those study and test items unsuccessfully recognized at final test. Both of these patterns are consistent with the above outlined target-context binding account, and furthermore, suggest the effect is not tied to a specific type of context memory (e.g., perceptual vs. associative).

Johnson (1992) outlines a framework which defines two types of events that can lead to strengthening the bind between target and context information in an episode: reinstatement and reactivation. When target information is re-presented to participants along with the relevant context information that was present during the original encoding event, reinstatement is thought to have taken place. However, one may recollect back to

[6]Although not concerned with source memory, Kornell et al. (2009) demonstrate a positive effect of unsuccessful retrieval on later item memory performance. However, fundamental differences between the present study and Kornell et al. suggest that the two sets of results are not incompatible. Of most relevance to the present study, Kornell et al. differed by requiring retrieval practice of previously unstudied information from semantic (rather than episodic) memory, providing feedback following each unsuccessful retrieval attempt, and using only semantically related paired associates. The latter two specifically (providing feedback and using semantically related materials) were critical in inducing memory benefits from unsuccessful retrieval. Considering these fundamental differences, there is little reason to expect the retrieval dynamics at play in Kornell et al. would apply to the current study.

the study episode in memory, thereby reactivating the original target in context. Importantly, reactivation is thought to enhance binding between context and target information (as is perceptual reinstatement of the relevant target and context features), and can occur in the absence of perceptually experiencing the information to-be-bound.

Applied to a testing effect design, participants can reactivate the contextual elements of the study episode during retrieval practice by recollecting (i.e., reactivating) the experience. In other words, the Johnson (1992) framework would suggest that testing allows one to mentally reinstantiate the original study episode. When *relative* estimates of recollection (i.e., not dependent on overall target recognition performance) are derived from the conditional remember response data in Experiment 1 for the UC, CS, and CF data sets, comparisons of each estimate between study and test conditions mirror context memory performance. For example, neither conditional remember responses nor context memory differs between study and CF test items. Yet, both the conditional proportion of remember responses and context memory performance are higher for UC and CS test items compared to study items. When the remember-know response data from Experiment 1 is extrapolated to the corresponding data set types in Experiment 2, the same pattern emerges. As such, the conditional remember response data, not the target recognition or recall data, were a more accurate predictor of context memory performance (at least on the aggregate), as would be expected by the target-context binding account as discussed above. In addition, it seems that *successful* retrieval practice, but not unsuccessful retrieval or additional study, is the mechanism that produces recollection enhancement.

While the current study has provided some evidence that testing enhances both item and context memory differentially, such conclusions must be tempered. Foremost,

experimental phases or tasks should not be thought of as being completely independent from each other. Hintzman (2011) argues against the process purity of, for example, encoding and retrieval occurring exclusively during study and test phases, respectively. In a similar vein, the current data do not suggest absolutely that the item and context memory (or remember and know responses, for that matter) tasks across experiments tapped into fundamentally and wholly different memory processes or information. Indeed, according to the binding account derived from Cook et al. (2006) above, "item information is an extraordinarily important mediator of being able to recover accurate source information" (Cook et al., 2006, p. 834).

An additional theoretical implication of the present study concerns the differentiation between the testing effect and generation effect. Often, generation and testing are conflated in the literature and in application (Karpicke & Zaromb, 2010; e.g., usually one engages in testing, not generation, if studying material that was previously learned in a class, as the material is tied to a specific episodic context). Mulligan (2004) and Mulligan et al. (2006), as discussed earlier, demonstrate evidence of either negative or null generation effects in context memory. However, the present study found the opposite: testing effects in both perceptual and associative context memory. These disparate results serve to illustrate fundamental differences between the generation and testing effects. Mulligan et al. (2006) argued that engaging in the act of retrieval drew processing away from the encoding of perceptual features during generation.

The necessity of conflating retrieval with initial encoding when examining the generation effect precludes one from observing whether retrieval in isolation may have any effect on episodic context memory. In this way, the present study and those of Mulligan (2004) and Mulligan et al. (2006) are not directly comparable. In the former,

the act of retrieval occurred following the presentation of to-be-tested context information, while the latter studies had both events occur simultaneously. Accordingly, Mulligan et al. (2006) show that engaging in retrieval alters what information is attended to during encoding (i.e., they demonstrate the consequences of retrieval on the concurrent *encoding* of context information, rather than the effect of retrieval on the *retention* of previously encoded context information).

This difference precludes the conclusion that episodic (testing) and semantic (generation) based retrieval have different effects on context memory. Follow-up research should look at the effects of testing on context memory when the relevant contextual information is presented during retrieval rather than initial encoding. In the case of the paradigm used in Experiment 1, this would mean introducing a font color manipulation at the intervening task, rather than initial study. Considering the frequent conflation of generation and testing in education, the present study suggests that retrieval practice can in fact be desirable, rather than detrimental, when the information to be learned is perceptual in nature (e.g., map or diagram learning).

Test-Induced Spillover Effects

A growing body of literature has found that retrieval practice can produce both target and non-target "spillover effects" in memory. Testing enhances transfer of learning (Butler, 2010; Johnson & Mayer, 2009; McDaniel et al., 2007; Roediger & Butler, 2011; Rohrer et al., 2010), which is critical for implementing retrieval practice in applied settings. Rarely is one required to retrieve previously learned information verbatim. Promoting encoding variability (i.e., encoding the same target information in different contexts or configurations) can be used to enhance transfer (Goode, Geraci, & Roediger, 2008), which can be induced through retrieval practice (e.g., Butler, 2010). Similarly,

engaging in thorough, deep retrieval practice can generalize the benefits of testing to untested but related material (e.g., Chan et al., 2006).

Common to these studies is the use of conceptually rich materials. For instance, Johnson and Mayer (2009) had participants learn information about lightning through an ecologically valid computer animation and narrative. Similarly, Butler (2010) promoted transfer and encoding variability with testing by providing participants with multiple practice tests on different variations of the same questions derived from prose passages. The current study extends the boundary of test-induced spillover effects through the use of unrelated single word lists (Exp. 1) and paired associates (Exp. 2). In each case, retention of untested but related context information was enhanced through testing, lending support to the validity of implementing findings derived from well controlled, but less ecologically valid testing studies to practical, real world settings. Furthermore, the present study may be conceptualized as demonstrating perceptual (Exp. 1) and associative (Exp. 2) flavors of RIFA, in that the act of retrieval (in part via enhancing recollection) not only facilitates retention of semantically related information (e.g., Chan et al., 2006; Cranney et al., 2009), but also perceptual and associative information surrounding the target.

Applications

A number of applications follow from the conclusions drawn from the present study. Regardless of the processes at play, the data suggest that engaging in retrieval practice can be a powerful method of boosting retention in educational environments given certain precautions. Experiment 2 saw a testing effect in target recall in the UC data, indicating that even only recalling approximately 50% of material during retrieval practice can pay off as a study technique, with the effect only growing as more

information is successfully retrieved. While Experiment 1 did not show an overall target testing effect in the UC data set, it is probable that the mediocre intervening test performance was culpable. Fortunately, in real world scenarios, we usually can devote as much time as needed to successfully retrieve information. For example, a student studying for an exam may want to practice recalling information through the use of flash cards. Even if the first retrieval practice attempt fails, nothing prevents the student from attempting retrieval again at a later time, say after reviewing the relevant information needed to answer the question.

While it has been established in the literature that implementing retrieval practice can have a practically significant effect even in real world scenarios (e.g., in the classroom, McDaniel et al., 2007), the more unique contribution from the current study stems from the effects of testing on context memory. In educational settings, information is rarely presented to students in isolation. Rather, the bulk of information to be learned is heavily inter-related. Test-induced retention advantages for both target and conceptually related information suggests testing can be a particularly potent learning tool in such settings. For example, a student may need to learn a set of historical facts and their associated dates and locations. Engaging in retrieval practice about the content of the event itself may also promote retention of the associated date or location as long as both pieces of information were initially studied during the same episode.

Similarly, enhancing perceptual context information can be equally important in educational settings. For example, a student may be attempting to learn a map of Africa. The results of Experiment 1 suggest that engaging in retrieval practice for the names of cities or countries may also facilitate learning of the locations of each city or country, despite the information not specifically being tested. Combined with recent evidence that

testing aids the *organization* of conceptually related information in memory (Zaromb & Roediger, 2010), testing can be a powerful learning enhancer for more than just the information as presented in an educational setting verbatim. Such is an important precondition to the adoption of testing as a learning mechanism in applied settings, as educators often often view tests with skepticism (Roediger & Karpicke, 2006a).

Concluding Remarks

The present study provides evidence of a retrieval-induced enhancement of recollection from a dual process perspective. The use of a remember-know task meshes well with such a perspective, as the measurements, assessed from the subjective, phenomenological experience of participants, map on to qualitatively different, independent processes (recollection and familiarity). However, the data are not incompatible from a single process viewpoint. Dunn (2004) provides strong evidence as to the compatibility of remember-know data with a single process framework. Similarly, apparent empirical dissociations (e.g., between target and context memory in the present study) do not necessitate dissociable processes (e.g., recollection and familiarity) to be explained (Benjamin, 2010). Thus, while the present study was framed within the context of dual process theory, it does not suggest an unrivaled interpretation of the data. Never the less, despite possible ambiguity in interpretation, the present study provides a foundation of evidence that should be built upon to further uncover the effects of retrieval practice on memory for contextual information.

References

Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2010). *Improving student learning with classroom quizzes: Three years of evidence from Columbia Middle Schoo*l. Poster presented at the 5th Annual Institute of Education Sciences Research Conference, National Harbor, MD.

Allen, G.A., Mahler, W.A., & Estes, W.K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8*, 463-470.

Asch, S.E., & Ebenholtz, S.M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society, 106*(2), 135-163.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research, 85*(2), 89-99.

Benjamin, A. S. (2010). Representational explanations of "process" dissociations in recognition: The DRYAD theory of aging and memory judgments. *Psychological Review, 117*(4), 1055-1079.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition 35*(2), 201-210.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition 15*(4), 657-668.

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

Brewer, G.A., Marsh, R.L., Meeks, J.T., & Clark-Foos, A. (2010). The effects of free recall testing on subsequent source memory. *Memory 18*(4), 385-393.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology:Applied 13*(4), 273-281.

Campbell, J., & Mayer, R.E. (2009). Questioning as an instructional method: Does it affect learning from lectures? A*pplied Cognitive Psychology, 23*, 747-759.

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

Carpenter, S.K., & DeLosh, E.L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19, 619-636.

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

Carpenter, S.K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474-478.

Carpenter, S.K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review 13*(5), 826-830.

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

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

Chan, J.C.K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49-57.

Chan, J.C.K., & McDermott, K.B. (2006). Remembering pragmatic inferences. *Applied Cognitive Psychology, 20*, 633-639.

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

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences *(second edition),* New Jersey: Lawrence Erlbaum Associates.

Cook, G.I., Marsh, R.L., & Hicks, J.L. (2006). Source memory in the absence of successful cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 828-835.

Cranney, J., Ahn, M.., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology, 21*(6), 919-940.

deWinstanley, P.A., Bjork, E.L., & Bjork, R.A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory, 4*(1), 31-48.

Dudukovik, N.M., & Knowlton, B.J. (2006). Remember-Know judgments and retrieval of contextual details. *Acta Psychologica, 122*, 160-173.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170-177.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 111*(2), 524-542.

Gardiner, J.M., Craik, F.I.M., & Bleasdale, F.A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1*(3), 213-116.

Gardiner, J.M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition, 7*, 1-26.

Geghman, K.D. & Multhaup, K.S. (2004). How generation affects source memory. *Memory & Cognition, 32*(5), 819-823.

Geraci, L., & McCabe, D. P. (2006). Examining the basis for illusory recollection: The role of remember/know instructions. *Psychonomic Bulletin & Review, 13*(3), 466-473.

Gillund, G. & Shiffrin, R.M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67

Glover, J.A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399.

Goode, M.K., Geraci, L., & Roediger, H.L., III. (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic Bulletin & Review, 15*, 662-666.

Hay, J.F., & Jacoby, L.L. (1996). Separating habit and recollection: Memory slips, process dissociations, and probability matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition,22*(6), 1323-1335.

Hintzman, D.L. (2011). Research strategy in the study of memory:Fads, fallacies, and the search for the “coordinates of truth.” *Perspectives on Psychological Science, 6*(3), 253-271.

Hinze, S.R., & Wiley, J. (2011) Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304.

Hockley, W.E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory & Cognition*, *27*(4), 657-664.

Jacoby, L.L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior, 22*, 485-508.

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13–47). Mahwah, NJ: Erlbaum.

Jang, Y., & Huber, D.E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition,34*(1), 112-127.

Johnson, C.I. & Mayer, R.E. (2009). A testing effect with multimedia learning. *Journal of Experimental Psychology, 101*(3), 621-629.

Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3-28.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30*(6) 823-840.

Kang, S.H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38*(8), 1009-1017.

Kang, S.H., McDermott, K.B., & Roediger, H.L., III. (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.

Karpicke, J.D., & Blunt, J.R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772-775.

Karpicke, J.D., & Roediger, H.L. III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition,33*(4), 704-719.

Karpicke, J.D. & Zaromb, F.M. (2010). Retrieval mode distinguished the testing effect from the generation effect. *Journal of Memory and Language, 62*, 227-239.

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

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. The American Journal of Psychology, 109(3), 451-464.

Kuo, T., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language, 36*, 188-201.

Madan, C.R., Glaholt, M.G., & Caplan, J.B. (2009). The influence of item properties on association-memory. *Journal of Memory and Language, 63*, 46-63.

Malmberg, K.J. (2008). Recognition memory: A review of the critical findings and an integrated theory for related them. *Cognitive Psychology, 57*, 335-384.

McCabe, D.P., & Geraci, L.D. (2009). The influence of instructions and terminology on the accuracy of remember-know judgments. *Consciousness and Cognition, 18*, 401-413.

McCabe, D.P., Roediger, H.L., III, & Karpicke, J.D. (2011). Automatic processing influences free recall: converging evidence from the process dissociation procedure and remember-know judgments. *Memory & Cognition, 39*, 389-402.

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., & Masson, M.E.J. (1985). Altering memory representations through memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(2), 371-385.

McDermott, K.B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34*(2), 261-267.

Meiser, T., & Sattler, C. (2007). Boundaries of the relation between conscious recollection and source memory for perceptual details. *Consciousness and Cognition 16*, 189-210.

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 609-622.

Morris, C.D., Bransford, J.D., & Franks, (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, *519-533.*

Mulligan, N.W. (2004). Generation and memory for contextual detail. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 838-855.

Mulligan, N.W., Beskin, M., & Peterson, D. (2010). Remember-know and source memory instructions can qualitatively change old-new recognition accuracy: The modality match effect in recognition memory.

Mulligan, N.W., Lozito, J.P., & Rosner, Z.A. (2004). Generation and context memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 836-846.

Mulligan, N.W., & Osborn, K. (2009). The modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 564-571.

Perfect, T.J., Mayes, A. R., Downes, J.J., & Van Eijk, R. (1996). Does context discriminate recollection from familiarity in recognition memory? *The Quarterly Journal of Experimental Psychology, 49A*(3), 797-813.

Pierce, B.H., & Gallo, D.A. (2011). Encoding modality can affect memory accuracy via retrieval orientation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 516-521.

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

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

Quamme, J.R., Yonelinas, A.P., & Norman, K.A. (2007). Effect of unitization on associative recognition in amnesia. *Hippocampus, 17*, 192-200.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition 21*(1), 89-102.

Roediger, H.L., III. (2000). Why retrieval is the key process in understanding human memory. In E. Tulving (Eds.), M*emory, consciousness, and the brain. The Tallinn conference* (52-75). Philadelphia: Psychology Press.

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

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

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

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233-239.

Runquist, W.N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11*(6), 641-650.

Runquist, W. N. (1986). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology 40*(1), 65-76.

Slamecka, N.J, & Katsaiti, L.T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26*, 589-607.

Sommer, T., Schoell, E., &Buchel, C. (2008). Associative symmetry of the memory for object-location associations as revealed by the testing effect. *Acta Psychologica 128*, 238-248.

Spencer, W.D., & Raz, N. (1995). Differential effects of aging on memory for content and context: A meta-analysis. *Psychology & Aging, 10*(4), 527-539.

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

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

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

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

Tulving, E. (1967). The effects of presentation and recall of materials in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*(1), 1-12.

Tulving, E., & Thomson, D.M., (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352-373.

Wartenweiler, D. (2011). Testing effect for visual-symbolic material: Enhancing the learning of Filipino children of low socio-economic status in the public school system. *The International Journal of Research and Review 6*(1), 74-93.

Wheeler, M.A., Ewers, M., & Buonanno, J.F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571-580.

Wheeler, M.A., & Roediger, H.L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results. *Psychological Science 3*(4), 240-245.

Whitten, W.B., & Bjork, R.A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Behavior, 16,* 465-578

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

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

Yonelinas, A.P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition 25*(6), 747-763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25(6), 1415–1434.

Yonelinas, A.P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

Yonelinas, A.P. & Jacoby, L.L. (1994). Dissociations of processes in recognition memory: Effects of interference and of response speed. *Canadian Journal of Experimental Psychology, 48*(4), 516-534.

Yonelinas, A.P., Kroll, N.E.A., Dobbins, I.G., & Soltani M. (1999) Recognition memory for faces: When familiarity supports associative recognition judgments. *Psychonomic Bulletin & Review, 6*(4), 654-661.

Zacks, R,T. (1969). Invariance of total learning time under different conditions of practice. *Journal of Experimental Psychology, 82*(3), 441-447.

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