THE INFLUENCE OF INTERVENING TASKS ON MEMORY

BY

KRISTIN M. DIVIS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Aaron S. Benjamin, Chair
Professor Kara D. Federmeier
Professor Brian H. Ross
Professor Daniel J. Simons
Professor Elizabeth Stine-Morrow

ABSTRACT

Learning new information is rarely done in isolation outside the laboratory. What one does *between* study sessions is important for later retrieval of the studied material. Previous research has shown that interleaved semantic retrieval leads to enhanced memory for later-learned items but to reduced memory for earlier-learned items (Divis & Benjamin, 2014). *Retrieval-induced distinction* provides one account for the opposite effects of retrieval on memory for early versus later material. By that view, retrieval serves to “spread out” the memory representations, making them (1) more distinct and less likely to interfere with each other and (2) rendering earlier events mentally more distant and harder to access at the time of test. Here, I expand on prior work examining the influence of intervening tasks on memory. Experiment 1 examines the effect of interleaved periods of task switching on memory for materials learned prior to or following the task switching intervention. Switching tasks (50% retrieval and 50% control) between periods of learning leads to similar benefits as those provided by a retrieval intervention. Experiments 2A and 2B examine the effects of intervening task on interlist intrusions. List distinction is enhanced when lists are separated by retrieval tasks, as seen by a reduction in interlist intrusions. Experiment 3 integrates and extends many of the components of prior studies by examining the influence of intervening retrieval on memory for complex, narrative text materials that share repeated, conflicting, and unique information. It highlights important boundaries to the influence of retrieval-induced distinction and provides a launching point for future studies examining the application of these intervening tasks in more ecologically relevant settings.

ACKNOWLEDGMENTS

This project would not have been possible without the input and support of many people. Thank you to my advisor, Aaron Benjamin, and committee members, Kara Federmeier, Brian Ross, Dan Simons, and Elizabeth Stine-Morrow for their continual support and feedback as I crafted this dissertation. Thanks also to the University of Illinois for its financial support of my research by awarding me the Illinois Distinguished Scholar Fellowship. Thank you to all the educators in my life who helped me to reach this milestone, from those in my hometown to the University of Nebraska to here at the University of Illinois. Your guidance and mentorship through critical periods of my life have been invaluable. And finally, thank you to my family and friends, who have always stood as a solid rock of love and encouragement. You pull me through the rough times and soar with me in the good ones.

TABLE OF CONTENTS

General Introduction

The ability to remember information is an essential skill in everyday life. For example, during final exams week, students are required to remember material learned throughout an entire academic semester across multiple courses. Drivers must remember where they last parked their cars and the best routes to their destinations. One is expected to remember the names of close friends, work colleagues, and newly met acquaintances. The need for and expectation of "good memory" is ubiquitous, yet memory is not perfect. Students forget information, drivers have difficulty finding their cars in parking lots, and names are forgotten at inopportune moments. Research lending insight into how memory can be enhanced (or hindered) creates the building blocks to understand and overcome the inherent fallibility of memory.

Two major sources of memory errors are interference from competing information (e.g., confusing yesterday's parking spot with today's) and a mismatch between one's state at test and initial encoding (e.g., having difficulty remembering the name of a work colleague when outside of the office). Memories that are similar to a target memory are more likely to impede successful retrieval of that target memory. Proactive interference occurs when memories acquired prior to the target memory cause interference; while retroactive interference occurs when memories acquired after the target memory cause interference. The greater the unique match between retrieval cues associated with the target memory and the memory query, the greater the chance of successful memory retrieval (e.g., Earhard, 1967; Roediger & Guynn, 1996; but see Nairne, 2002). On a similar note, the greater the match between the context at retrieval and encoding, the more likely one is to successfully retrieve the target memory (e.g., Estes, 1955).

Here I will narrow my focus to the impact of *intervening tasks* on interference and memory. Very rarely outside of a controlled laboratory setting is information learned and tested in isolation from other events. For example, the typical college student attends multiple classes each day, separated by events such as walking across campus, talking with friends, studying notes, or eating lunch. Even common laboratory experiments have “filler” tasks around and within the main tasks of interest. While the influence of strategies such as spacing (e.g., Ebbinghaus, 1964) and testing (e.g., Roediger & Karpicke, 2006) have been studied in depth, the influence of intervening tasks on memory has yet to be fully explored. As I will demonstrate here, what one does *between* studying matters for later memory performance. Specifically, I will focus primarily on the influence of intervening tasks involving *unrelated*, *semantic retrieval* on memory for information prior to and after the intervening tasks.

My primary focus is understanding and expanding on *what* occurs due to intervening retrieval. I will first focus on prior work examining the influence of both episodic and semantic retrieval. In the subsequent section, I will map out theoretical reasons for *why* these effects occur. The explanation most consistent with empirical evidence is that interleaved retrieval leads to more distinct memory representations, but alternative explanations are also considered. Experiment 1 tests whether task switching can drive the retrieval effects seen in previous studies. Experiments 2A and 2B expand on the impact of intervening retrieval, confirming that it leads to reduced confusions between retrieval-segregated material. Experiment 3 applies interleaved retrieval to more complex, overlapping materials.

Episodic Retrieval

Episodic memory is for specific events or experiences, whereas semantic memory is for acquired facts, meanings, or concepts about the world. Episodic retrieval of previously studied information is known to influence later memory. The most relevant, well-known result is the *testing effect*—information that is tested is better remembered at a later time than information that is only restudied (for a review, see Roediger and Karpicke, 2006). However, the influence of episodic retrieval extends beyond memory for the information that is actually tested; it also affects memory for untested information learned prior to and after the episodic retrieval event.

Prior Memory

Episodic retrieval influences memory for material prior to the retrieval event—even when that material was not tested. Shiffrin (1970) first examined this influence using the "list-before-last" paradigm. Participants continually studied lists of words. After each list, they were tested on the list *prior* to the previously studied list. For example, after studying List 3, participants were tested on List 2. Shiffrin found that while the length of the *target* list (i.e., List 2 in the previous example) influenced memory performance for that list (the list length effect; Murdock, 1960; Roberts, 1972), the length of the *intervening* list (i.e., List 3 in the previous example) did not. Follow-up studies showed that this effect only occurred when retrieval (as opposed to an unfilled pause) occurred between the lists: in the absence of the intervening retrieval task, longer intervening lists negatively influenced memory performance for the target list (Jang & Huber, 2008). While not emphasized by the authors, overall performance also tended to be lower when episodic retrieval separated the tasks as opposed to a pause (Jang & Huber, 2008), an effect consistent with later work (e.g., Divis & Benjamin, 2014).

Abbreviated versions of the list-before-last paradigm using only three lists replicated and extended these findings. When compared to a math control condition (Sahakyan & Hendricks, 2012) or restudy control condition (Sahakyan & Smith, 2013), intervening retrieval led to overall *lower* performance on the target list and *reduced* intrusions from the intervening list.

Manipulating the difficulty of the intervening retrieval task did *not* appear to influence either correct target list recall or intrusions from the intervening list (Sahakyan & Hendricks, 2012).

Future Memory

Episodic memory retrieval also impacts future memory (beyond the testing effect). Tulving and Watkins (1974) originally studied memory interference in a paired cued recall paradigm (AB-AC). When participants were tested on the AB pair before studying the AC pair, eventual memory for the AC pair was higher (compared to when the AB pair was not tested). This finding is in line with other work showing that items from untested lists are more likely to intrude on recall than items from tested lists (Darley & Murdock, 1971), suggesting that the enhanced performance of AC items in the Tulving and Watkins (1974) experiments were spurred by reduced intrusions from the AB items. Later research directly addressed this question in a multilist learning paradigm by examining the buildup of proactive interference across lists when previous lists were or were not tested (Szpunar, McDermott, & Roediger, 2008). In Experiments 1A, 1B, and 3 of that paper, participants studied five lists of words, separated by either a free recall test of the previous list, restudy of the previous list, or a control math task. In line with previous research, a final test of List 5 showed enhanced performance for participants in the test condition compared to either of the other two conditions. Notably, those in the test condition were also less likely to intrude items from Lists 1-4 onto recall of List 5, indicating a reduction in proactive interference. Experiment 2 directly examined the buildup of proactive interference by testing only one of the lists during the study block and seeing how levels of proactive interference changed based on the number of untested lists before the final test of List 5. Proactive interference steadily increased with more untested lists, especially for the first few lists. Reduction in proactive interference appears to be a major factor in the benefit of episodic retrieval on memory for future material.

Semantic Retrieval

Studying the influence of episodic retrieval on memory has some limitations. Episodic retrieval almost always involves retrieval of information learned in the current experiment. For example, memory for later items is enhanced when prior information is retrieved (e.g., retrieving Lists 1-4 after study led to better memory for List 5 items; Szpunar et al., 2008). That benefit could be driven by an increase in the memory strength of the retrieved information, making it better defined and thus less likely to be incorrectly identified as target list information (and therefore less likely to “muddle” recall of target list items). An alternative explanation is that something about the process of retrieval itself—absent of any direct ties to potentially interfering information from other lists—enhances memory for future items. Pastötter and colleagues (2011) delved into this question by expanding Szpunar et al.’s (2008) multilist learning paradigm to include intervening tasks of semantic retrieval (generate items from a given category), retrieval from short-term memory (2-back task), or math (count backward by 3s) in addition to episodic retrieval (recall the previous list) or restudy of the previous list. Memory performance for the final list was enhanced in the retrieval conditions (compared to the distractor or restudy conditions), *regardless of the type of retrieval*. Furthermore, intrusions from previous lists were reduced in both the episodic and semantic retrieval conditions. One experiment in the list-before-last literature (Jang & Huber, 2008) also utilized a semantic retrieval task, albeit one much different than the semantic generation task used above. An intervening task involving lexical completion (e.g., indicate whether an “i” or “e” should be included to make an item a word) led to similar list isolation effects as episodic free recall. These results suggest there may be something about the process of retrieval itself (not just retrieving potentially competing information) that enhances future memory. They also indicate that in multilist learning paradigms, semantic retrieval tasks may be utilized instead of the more common episodic retrieval. Semantic retrieval (as opposed to episodic retrieval) has the benefit of avoiding the confounds associated with testing material studied during the lab session.

Divis and Benjamin (2014) further explored the effect of interleaved retrieval, studying its influence on both prior and future information in one procedure using semantic retrieval (see Figure 1 for a summary of these results). In Experiment 1, participants studied five lists of unrelated words, each separated by either a distractor counting task or semantic retrieval generation task (as in Pastötter et al., 2011). They then completed a free recall test for either List 1 or List 5. Memory was *enhanced* for List 5 but *reduced* for List 1 in the retrieval condition compared to the distractor condition. Furthermore, prior list intrusions on List 5 were reduced in the retrieval condition compared to the distractor condition; interlist intrusions on List 1 were at floor for both conditions. These findings replicate and combine the effects of episodic retrieval on prior and future memory discussed above, using semantic retrieval. Once again, these results suggest there is something about the act of retrieving—and not just retrieving potentially competing information—that influences memory for surrounding items. A reduction in proactive interference was found yet again; retroactive interference did not appear to be a player in this design (as seen from the low level of intrusions from Lists 2-5 on List 1, regardless of condition). Experiment 2 extended the effect to four narrative text materials (as opposed to five lists of unrelated words). Once again, memory for later material (Text 4) was benefitted but memory for earlier material (Text 1) was hindered when the texts were separated by a semantic generation task compared to a distractor counting task. Interleaved semantic retrieval reliably led to reduced memory performance for earlier material but increased performance for later material.

Retrieval—whether episodic or semantic—has the consistent effect of benefiting memory for later items but hindering memory for prior information. In the next section, I will discuss reasons *why* this effect may be happening, with an emphasis on retrieval-induced distinction.

Retrieval-Induced Distinction

Any explanation for the effects of intervening retrieval must be able to simultaneously account for reduced memory for pre-retrieval items and increased memory for post-retrieval items. On explanation that fits well with the current body of evidence is that retrieval serves to spread out the memory representations, making them more distinct from one another. See Figure 2 for a diagram of this *retrieval-induced distinction*. Retrieval-induced distinction has two major consequences: (1) Material prior to the retrieval event becomes more different and "farther back" in mind as the memory representations spread out, and (2) Memory representations that are more distinct from one another have fewer overlapping elements and therefore should be less likely to interfere with one another.

Retrieval-induced distinction provides a framework for both understanding the known effects of intervening retrieval within a multilist learning paradigm and predicting new ones. As Figure 2 shows, the spreading out of memory representations for the lists via retrieval leads to the List 1 representation being farther back and more difficult to access at the time of the criterion test. That increased mental distance predicts poorer memory performance on List 1 items when the lists are separated by retrieval tasks. Previous research using both episodic and semantic retrieval support this prediction (e.g., Jang & Huber, 2008; Divis & Benjamin, 2014).

Lists separated by a retrieval event should also be less likely to interfere with one another because they share fewer overlapping memory cues. Research using both episodic and semantic retrieval as the intervening task found the expected reduction in proactive interference (e.g., Szpunar et al., 2008; Pastötter et al., 2011; Divis & Benjamin, 2014). This reduction in proactive interference accounts for the memory benefit for materials that come after the retrieval tasks (e.g., List 5).[1]

Retrieval-induced distinction via context change

One possible mechanism for retrieval-induced distinction is that retrieval leads to changes in the contextual cues associated with each memory representation. During encoding, fluctuating contextual cues[2] (drawn from an internal mental context) are bound to the memory traces for the encoded information (e.g., Stimulus Sampling Theory, Estes, 1955; Mensink & Raaijmakers, 1988). The greater the match between the context at encoding and the context at time of retrieval, the more likely the target memory will be successfully retrieved (e.g., Estes, 1955; Kahana & Howard, 2005; Tulving & Thompson, 1973). Retrieval-induced distinction may be driven by retrieval leading to more internal context change than would normally occur

[1] The existing work on the memory effects of intervening retrieval tasks has primarily used free recall tasks. Free recall tasks tend not to elicit large amounts of interference. In fact, while previous work found relatively small but significant amounts of proactive interference using a free recall task, the retroactive interference was negligible (e.g., Divis & Benjamin, 2014). In those designs, the proactive interference was beneficially reduced by intervening retrieval but retroactive interference was at floor, regardless of condition. If the experimental design was changed so that retroactive interference was more likely to occur, one would expect intervening retrieval to also lead to a beneficial reduction in retroactive interference. In that case, the reduction in retroactive interference would *benefit* memory for pre-retrieval information, potentially counteracting the negative effects of the memory representation being more distant in mind. Experiments 2A and 2B in the current manuscript take the first steps toward testing these predictions.

[2] Here, context refers to anything in the current internal state. That can range from conscious thoughts to very basic biological processes. The theoretical interpretation provided here is agnostic to the exact content of the context. What matters is that there are many internal states occurring, all fluctuating at different rates. Retrieval serves to speed up those rates of fluctuation, making the contextual cues associated with material prior to and after the retrieval event more disparate.

otherwise (i.e., the “spreading out” of memory representations is due to more disparate contextual cues). Mental context change has previously been used to explain similar memory effects (e.g., in directed forgetting).

Context change in directed forgetting. The implications of a context change perspective have been most closely studied in directed forgetting studies. In the standard list-method directed forgetting paradigm, participants study a list of words, are told to either forget or remember that list, and then study a second list of words. Performance on a subsequent memory test reveals enhanced memory for List 1 items but *reduced* memory for List 2 items for participants in the “remember” compared to “forget” condition (e.g., Reitman, Malin, Bjork, & Higman, 1979). Thus, the forget cue not only leads to the intuitive result of poorer memory for the to-be-forgotten list, but it also *enhances* memory for the subsequent list. One explanation for this benefit is that being allowed to forget List 1 leads to less proactive interference from List 1 onto List 2 (Bjork & Bjork, 1996).

Effects similar to those of a forget cue are also found when internal context is manipulated away from the current state. Imagination tasks (e.g., imagining being invisible, walking through one’s childhood home, or going on vacation) interleaved between two lists of items also lead to reduced performance on the earlier list and enhanced performance on the later list (Sahakyan & Kelley, 2002; Delaney et al., 2010). Notably, effects of a larger magnitude are found when the imagination task is temporally or spatially farther away (e.g., domestic versus international vacation; Delaney et al., 2010). Unlike imagination tasks, activities such as solving math problems (Sahakyan & Kelley, 2002), counting tasks (Pastötter & Bäuml, 2007; Pastötter et al., 2011; Sahakyan, Delaney, & Goodman, 2008), number searches (Mulji & Bodner, 2010), and speeded reading tasks (Delaney et al., 2010) do *not* mimic the directed forgetting effect. One of the putative differences between these two sets of tasks is the amount of mental context change involved. A theoretical perspective emphasizing context change has been a useful tool in the field of directed forgetting. Context change (and the knowledge gained from the directed forgetting paradigms) could also be helpful in understanding other memory effects (such as the influence of retrieval).

Context change by retrieval. While the influence of context change has been most closely studied in the directed forgetting literature, an emergent field of study links increased context change to retrieval tasks as well. *Imagining* the events used in the directed forgetting-like paradigms outlined above (e.g., Sahakyan & Kelley, 2002) may rely on some of the same mechanisms as intentionally *retrieving* those events (e.g., Schacter, Addis, & Buckner, 2008).

Divis & Benjamin (2014) proposed that retrieving causes the rate of context fluctuation to increase[3], creating a greater disparity between the contexts prior to and after the retrieval event than would have otherwise occurred. Contextual cues bound to the memory for information pre- and post-retrieval will thus be more dissimilar than if no retrieval event happened. Similar ideas about the role of context have been proposed within multilist learning paradigms (e.g., Jang & Huber, 2008; Pastötter et al., 2011; Sahakyan & Hendricks, 2012) and the directed forgetting effect (e.g., Sahakyan & Kelley, 2002; Delaney, Sahakyan, Kelley, & Zimmerman, 2010). The spreading out of memory representations found in retrieval-induced distinction could be primarily driven by retrieval leading to more distinct contextual cues being tied to those memory representations at the time of encoding.

Context-based explanations can further inform the mechanisms of the two consequences of retrieval-induced distinction (pre-retrieval items becoming farther back in mind and reduced

[3] Contextual elements are continuously fluctuating in and out (i.e., Estes' Stimulus Sampling Theory, 1955). This interpretation posits that retrieval increases that rate of fluctuation, so that the contextual elements cycle out more quickly than would normally occur without the retrieval event.

interference between retrieval-segregated items). First, each retrieval event leads to a shift in the rate of context change and pushes one further down the “contextual stream.” This leads to a greater mismatch between contexts prior to and after retrieval. Importantly within a multilist learning paradigm, this predicts a greater disparity between the contexts at the beginning of the study session (e.g., List 1) and the criterion test at the end of experiment for those who complete interleaved retrieval tasks compared to those who do not (see Figure 2). The greater the disparity between the study and test contexts, the less likely one will successfully retrieve the studied information. Second, lists that are separated by a retrieval event become more contextually distinct. When two lists are more contextually distinct, they will have fewer overlapping retrieval cues and will be less likely to interfere with each other (either proactively or retroactively).[4]

Potential limitations of context change theories. A context change interpretation of retrieval-induced distinction accounts well for the opposite effects of intervening retrieval on prior and future memory. However, an inherent issue when studying the influence of context change on memory is being able to operationalize it, independent of the memory itself. Recent work has shown that manipulations thought to influence context change not only affect memory but also perceived time estimates (Sahakyan & Smith, 2013). The more context change, the longer the perceived elapsed time. Furthermore, theories based on context change have allowed for accurate predictions in novel paradigms (e.g., Divis & Benjamin, 2014). The growing body of evidence suggests that context change is a useful explanatory construct, but it will be

[4] Divis & Benjamin (2014) created a mathematical model that simulated the effects of increased context fluctuation. Their simulations fit the predictions of overall poorer memory for earlier items and better memory later items quite well. Additionally, the model made the novel prediction that a sufficient delay in the criterion test will wash out the differences introduced by the interleaved retrieval manipulations. When the test context is shifted far enough away from the study session context (by a long delay), the differences between the list contexts within the study session will be *relatively* smaller compared to the criterion test context than if no delay had occurred. As predicted, delaying the criterion test alleviated the effect of interleaved retrieval on memory performance (Experiment 3).

strengthened by further research highlighting both novel predictions and ruling out competing hypotheses.

Alternative Explanations

The challenge in interpreting the effects of interleaved retrieval within a multilist learning paradigm is accounting for *both* the reduced memory for prior material and enhanced memory for later material. Any number of theories can account for one or the other result; for example, the act of *practicing* retrieval during the interleaved tasks might lead to superior retrieval later by a simple practice effect or by clarifying the nature of the upcoming test (Finley & Benjamin, 2012). That simple account could explain the better memory for later material but not the poorer memory for earlier material.

An alternative explanation to retrieval-induced distinction is that these effects arise as a consequence of inhibition[5]. Retrieval could lead to inhibition of previous material, thus hindering one's ability to later successfully access that material and leading to overall lower memory performance on information prior to the retrieval event. When an earlier memory is inhibited, it will also be less likely to proactively interfere (thus leading to increased memory for material following the retrieval event). A prediction of this hypothesis is that modifying the Divis and Benjamin (2014) paradigm to also include a retrieval event *after* List 5 would lead to a reduction in List 5 performance compared to if there was not a retrieval event. However, a theoretical understanding based on retrieval-induced distinction would make the same prediction as a consequence of retrieval spreading out the mental distance between List 5 and the criterion test.

[5] Here, I will assume inhibition refers to inhibition of the prior material itself. Another option is that it is the *context* associated with earlier material that is being inhibited. Whether context is being inhibited or the rate of context fluctuation is being changed is very difficult to differentiate. Currently, no evidence exists to distinguish the two (very similar) perspectives.

If a retrieval event was added after List 5, the overlap between List 5 and the criterion test would be reduced (which also predicts a reduction in performance). In order to differentiate the predictions of these two views, one would first have to control for the relative mental distance (i.e., placement within the contextual stream) between the to-be-tested list and criterion test (e.g., by having an equal total number of retrieval events between the two[6]).

Similar to the inhibition hypothesis, retrieval could also be serving to interrupt consolidation of prior material. If the information is not consolidated as well, it will be more difficult to successfully retrieve that information at a later time. Poorly consolidated information should also be less likely to proactively interfere with later material. This hypothesis suggests that material at the end of a list prior to a retrieval event will be most susceptible to the negative consequences of retrieval.

The biggest weakness of both the inhibition and interruption of consolidation hypotheses is their inability to simultaneously explain the episodic and semantic retrieval findings. Both interleaved episodic and semantic retrieval lead to enhanced performance on later, untested material within a multilist learning paradigm. Pastötter and colleagues (2011) showed that the magnitude of the effects of overall memory performance and reduction in proactive interference are similar, regardless of retrieval type. An interpretation based on retrieval-induced distinction would attribute this enhancement of later memory to the reduced proactive interference that arises from more effective list isolation. Perspectives that involve inhibition and interruption of consolidation predict that same enhancement due to a reduction in proactive interference. However, the reduction in proactive interference occurs due to a weakening of prior material. Although such a mechanism is plausible for interleaved semantic retrieval, episodic retrieval of

[6] The design of Experiment 2A in the current manuscript meets this criterion for Lists 1 and 3 relative to the final test. An equal amount of retrieval events occur between those lists and the final test, regardless of condition.

prior material serves to *enhance* retrieved material—not weaken it (i.e., the testing effect). Perhaps semantic retrieval serves to weaken prior material and thus reduce proactive interference, while episodic retrieval serves to enhance prior material, leading to less source confusion (and thus less proactive interference) at the time of test. However, that explanation is not satisfying, especially given the similar magnitude of effects shown by semantic and episodic retrieval.

Introduction to Experiments

The current set of experiments further explores the impact of intervening tasks on memory. Experiment 1 focuses on the effect of task switching. Experiments 2A, 2B, and 3 examine how intervening retrieval affects interlist interference and facilitation for both basic lists of words and more complex text materials.

In the typical multilist design, the intervening retrieval task is preceded by a brief portion of the control task (e.g., Pastötter et al., 2011; Divis & Benjamin, 2014). The purpose of such a design is to control for the transition from study events to intervening events, but that control comes at a price: participants effectively complete two tasks (the control task and a retrieval task) in the retrieval condition but just one in the control condition. No study to date has been designed to specifically test the influence of such task switching on memory. Perhaps the list distinction effects have been driven by that task switching rather than retrieval itself. Experiment 1 compares a pure distractor interleaved task, a pure retrieval interleaved task, and a hybrid interleaved task (composed of half of each) to directly address the potential confound of task switching and to determine the influence of task switching in general.

The free recall tests used in previous research (e.g., Divis & Benjamin, 2014) tend to elicit few interlist intrusions. Reduced intrusions due to enhanced list distinction are a key component to the underlying framework used to explain the effects of intervening retrieval, so it is critical to find a design in which intrusions are more common and the predicted reduction in intrusions can be tested. Experiments 2A and 2B utilize a new test paradigm in which intrusions should be more frequent. These experiments are specifically designed to lend insight into the level of segregation between lists that are separated by an intervening retrieval or control task.

Finally, Experiment 3 was designed to (1) consolidate the effects found in earlier individual experiments into one large study and (2) see if those effects transfer to more complex, related narrative materials (eye witness testimonies about the same crime).

Experiment 1: Switching

Experiment 1 explores the influence of switching tasks in addition to the effect of retrieval seen in earlier work. Which tasks do and do not create increased distinction between materials has not been thoroughly mapped out. Switching from one type of a task to another type of task might also lead to more distinct memory representations, which is important for the retrieval-based interpretation of prior work.

Earlier work examining the influence of retrieval (Pastötter et al., 2011; Divis & Benjamin, 2014) employed a procedure in which the retrieval task was preceded by a distractor control task (e.g., [Study List 1 – 30 seconds distractor – 60 seconds retrieval – Study List 2, …] compared to [Study List 1 – 90 seconds distractor – Study List 2, …]). Potentially, it is the *switching* of tasks between studying lists and not retrieval that drives the effects seen in those studies (or alternatively, some combination of switching and retrieval). Previous work suggests that merely switching tasks is not sufficient to explain these effects, at least for future memory. Pastötter et al. (2011) included distractor counting, restudy (preceded by counting), episodic retrieval (preceded by counting), semantic retrieval (preceded by counting), and short term memory (preceded by counting) intervening tasks. They found a benefit for List 5 performance for the three retrieval conditions compared to the distractor and restudy conditions. Even though restudying was preceded by 30 seconds of counting, it led to the same effect as spending the entire time counting. However, the effect of task switching has not been directly or intentionally tested for its influence on either earlier lists (List 1) or later lists (List 5).

Experiment 1 examines the effect of both intervening semantic retrieval and task switching on memory for earlier and later material. Here, task switching refers to switching tasks within a single intervening session (i.e., completing two different tasks between studying lists). See Figure 3 for the experimental design.

Method

Participants

One hundred sixty-eight undergraduates at the University of Illinois at Urbana-Champaign completed this experiment for course credit. The data from 32 participants was dropped for failing to properly follow instructions (they typed at least one word from a previously studied list when they were supposed to type in words belonging to the given category in the semantic retrieval task), leaving a total of 136 participants' data for analysis.

Design

Type of task (pure semantic retrieval, pure distractor, semantic retrieval followed by distractor, or distractor followed by semantic retrieval) and list tested (List 1 or List 5) were manipulated between subjects, for a total of eight conditions. Memory performance was measured via a free recall test.

Materials

Fifty words (average word length = 5.10 letters, *SD* = 1.42 letters) were drawn from the University of South Florida Free Association Norms (Nelson, McElvoy, & Schrieber, 1998) to create five study lists of ten words each. Words were randomly assigned to each list, did not belong to the semantic retrieval categories, and were not related to one another. No words were repeated in the experiment.

Procedure

Participants completed the experiment individually in separate rooms. PC-style computers programmed using MATLAB with the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997) were used to present stimuli and record responses in the experiment. Prior to beginning the study phase, participants were presented with instructions informing them that they would be studying lists of words and that their memory for those words (and which lists they belonged to) would be tested at the end of the experiment (an example was given). They were also told that the lists would be broken up by different tasks.

During the study session, participants studied five lists of words (labeled numerically), separated by intervening tasks. Words were presented one at a time for 4 seconds each, with an interstimulus interval of 500 ms. The intervening tasks consisted of a combination of the distractor and retrieval tasks, depending on the experimental condition. In the distractor task, subjects counted aloud backward by 3s from a randomly generated 3-digit number for 45 seconds. The number changed with each iteration of the distractor task. In the retrieval task, participants listed exemplars from a given semantic category (vegetables, 4-legged animals, sports, or professions)[7]. Categories were never repeated across iterations of the retrieval task and the order was randomly determined. In the pure distractor condition, participants counted backward from two numbers (for a total of 90 seconds). In the pure retrieval condition, participants listed exemplars from the same category for 90 seconds. In the switch conditions, participants spent 45 seconds doing each type of task (both distractor and retrieval) for a total of 90 seconds. Those in the distract-retrieve condition started with the distractor counting task, while those in the retrieve-distract condition started with the retrieval task. Participants

[7] These tasks cannot easily be equated on relative difficulty across all dimensions. However, Sahakyan & Hendricks (2012) showed that relative difficulty of the retrieval task did not lead to different magnitude of effects, at least within the similar list-before-last paradigm.

completed the *same* type of intervening task across the experiment (e.g., if they were in the pure retrieval condition, they performed the pure retrieval task after lists 1-4).

After finishing studying List 5, participants spent 60 seconds completing simple 2-digit addition problems. Memory was then tested via free recall for either List 1 or List 5, depending on condition. Participants typed in all the words they could remember from their assigned list. No time limit was imposed.

Results and discussion

In general, the *a priori* analyses of interest were the simple effects of task on each list for proportion correct and interlist intrusions. More specifically, I planned to compare (1) pure retrieval versus pure distractor, (2) pure distractor versus switching, and (3) pure retrieval versus switching. Initially, the two subdivisions of the "switching" condition were created as a counterbalancing measure and were assumed to be the same. That assumption appeared to be premature for List 5 performance correct measures; *post hoc* analyses were also conducted to examine that potential difference.

Significance levels for all statistical tests were set at an $\alpha < .05$ level. Random effects models fit using maximum likelihood estimation and Satterthwaite approximations to degrees of freedom in the lme4 package (Bates, Maechler, & Dae, 2011) in R software (R Development Core Team, 2008) were used to analyze the intrusion data. Accuracy (binomial) data was fit using Laplace estimation. See Appendix A for model fitting details; see Appendix B for Bayes factors. Proportion correct as a function of list tested and intervening task are shown in Figure 4a. Average number of interlist intrusions (i.e., intrusions from words from the other lists onto the tested list) are shown in Figure 4b. Notably, inferential analysis of the intrusion data was limited by the low occurrence of interlist intrusions (as expected from previous work; e.g., Divis & Benjamin, 2014).

Proportion Correct

The best fit model for proportion correct included fixed effects for intervening task (pure retrieval, pure distractor, or switch) and list tested (List 1 or List 5), along with random intercepts for words.

<u>Simple effects of task for List 1.</u> Performance on List 1 was significantly higher in the pure distractor condition compared to both the pure retrieval (z = 2.12, p = .034) and switch (z = 3.569, p <.001) conditions. Participants in the pure retrieval and switch conditions did not differ in List 1 memory performance.

<u>Simple effects of task for List 5.</u> Performance on List 5 was significantly lower in the pure distractor condition compared to both the pure retrieval (z = 5.672, p < .001) and switch (z = 4.75, p < .001) conditions. While proportion correct for List 5 was numerically higher in the pure retrieval compared to switch condition, that difference was not reliable (z = 1.803, p = .073).

Interleaved pure retrieval led to enhanced memory for later items but reduced memory for earlier items, compared to the control distractor condition. These results replicate previous findings (Divis & Benjamin, 2014), controlling for switching task. The switch condition (which was half retrieval and half distractor tasks) led to the same pattern of results relative to the distractor control condition as the pure retrieval condition.

While null effects should be interpreted with caution[8], no reliable differences were found between the pure retrieval and switch conditions. This suggests that (1) it is not the *amount* of

[8] Although see the Bayes factors reported in Table 3, indicating weak support for the null hypothesis.

retrieval that matters (e.g., 90 versus 45 seconds) but that some retrieval occurs; or (2) switching tasks combined with a small amount of retrieval lead to similar magnitude of effects as extended retrieval. Notably, performance correct was consistently *numerically* higher in the retrieval compared to switch conditions, suggesting a more complex relationship is occurring than "amount of retrieval".[9]

Interlist Intrusions

The best fit model for interlist intrusions included fixed effects for intervening task (pure retrieval, pure distractor, or switch) and list tested (List 1 or List 5), along with random intercepts for words.

<u>Simple effects of task for List 1.</u> The model revealed no significant differences in number of interlist intrusions onto List 1 when comparing pure distractor, pure retrieval, and switch conditions.

<u>Simple effects of task for List 5.</u> Fewer interlist intrusions occurred onto List 5 in both the pure retrieval ($t(136) = 2.884$, $p = .005$) and switch ($t(136) = 3.331$, $p = .001$) conditions relative to the pure distractor condition. The model revealed no significant difference in interlist intrusions onto List 5 between pure retrieval and switch conditions.

Intrusion analysis was limited due to the overall low occurrence, a limitation particularly prominent for List 1. Once again, these results replicate the pattern of results shown in previous work (Divis & Benjamin, 2014), controlled for switching tasks. Retrieval (and switching) led to reduced interlist intrusions for List 5 (i.e., a reduction in proactive interference) and a null result

[9] The fact that amount of retrieval is unable to fully explain the effects of interleaved retrieval is consistent with work in the list-before-last paradigm showing that the difficulty of the retrieval task also does not influence the magnitude of the effect (Sahakyan & Hendricks, 2012).

for List 1 (but the intrusions were at floor, indicating the design elicited very little measurable retroactive interference).

Exploratory analyses

No difference in switch counterbalancing conditions (distract-retrieve or retrieve-distract) was originally expected. However, the List 5 results (as seen in Figure 4a) suggest that something more complex may be going on. The effects in the retrieve-distract condition appear to be of similar magnitude to those seen in the pure retrieval condition, with the distract-retrieve condition falling somewhere between those and the results from the pure distractor condition. This same divergence is not seen in the List 1 results nor in the interlist intrusion data.

To evaluate whether the apparent difference between the distract-retrieve and retrieve-distract conditions was more than sampling error, a new model was run that included the two counterbalancing variants of the switch condition. The best fit model included the fixed effect of task (pure retrieval, pure distractor, retrieve-distract, or distract-retrieve) and list tested (List 1 or List 5), along with random intercepts for words. The model revealed a significant difference between retrieve-distract and distract-retrieve conditions for List 5 performance (z = 3.018, p = .003), with higher performance in the retrieve-distract condition compared to the distract-retrieve condition. One interpretation of these findings is that the magnitude of the gap between studying the previous list and semantic retrieval is important. The closer the retrieval to the previous list, the greater the benefit to later memory. Notably this does not appear to be related to proactive interference—the same pattern of results was not seen in the interlist intrusion data.

Summary

Experiment 1 provided a replication of earlier work (Divis & Benjamin, 2014): interleaved pure semantic retrieval led to reduced memory performance on earlier material and enhanced memory performance on later material compared to a pure distractor task. Importantly, this occurred even with complete control over the amount of switching (i.e., by eliminating the short block of the distractor task before beginning the retrieval task). Experiment 1 also showed that task switching leads to a similar enhancement for later material but hindrance for earlier material relative to a distractor control task. Notably, the switch condition included a period of retrieval as well.

Merely switching tasks is not enough to explain the effects of interleaved retrieval. However, whether completing multiple tasks in the switch condition or just completing a brief amount of retrieval (or even some combination of the two) led to the benefits and deficits seen in the switch condition remains unclear. While performance was numerically higher in the retrieval compared to switch condition for List 1 material, that difference was not reliable; and pure retrieval and retrieval combined with the distractor task led to similar results on List 1, regardless of order.

Pure retrieval and retrieval *followed by* distractor counting led to virtually equivalent performance on List 5; however, those who completed the distractor counting *prior to* the retrieval task had overall lower performance than those who did the retrieval task first. Order of retrieval when switching tasks matters for List 5 performance. Notably, interlist intrusions onto List 5 are reduced in the pure retrieval and both switch condition orders relative to the distractor task. The discrepancy in switch condition performance only manifests in accuracy. None of the theoretical interpretations considered here can easily account for these findings. While a perspective that proposes an interruption in consolidation is in line with the fact that retrieval leads to a bigger enhancement in memory for later material when it is closer to the prior material (as in the pure retrieval and retrieval-distractor conditions), that perspective would also predict that a similar pattern should show up in the interlist intrusions. Further work would need to be done—including validating through replication that it is not a spurious effect—in order to more fully understand the underlying mechanisms of this discrepancy. However, Experiment 1 demonstrates that *both* pure retrieval and switching tasks (one of which includes retrieval) consistently lead to enhanced memory for later items and reduced memory for earlier items compared to a control distractor task.

Since the retrieval manipulation used in earlier work (e.g., Pastötter et al., 2011; Divis & Benjamin, 2014) survived Experiment 1 intact, further research can test critical predictions and boundaries to the effect of interleaved retrieval. Next, I narrowed in on the impact of retrieval on interlist intrusions.

Experiments 2A and 2B: Intrusions

The main goals of Experiments 2A and 2B were to (1) more closely examine the pattern of interlist intrusions evoked by interleaved semantic retrieval and (2) use a more efficient within-subjects design for exploring the effects of interleaved semantic retrieval[10]. Here, the intervening task was alternated between studied lists (see Figure 5). Lists separated by a semantic retrieval task should be more distinct from one another (and therefore less likely to intrude on one another) compared to lists separated by the distractor task. Instead of a free recall test (which leads to a limited number of interlist intrusions, as seen in Experiment 1 in the current manuscript and Divis & Benjamin, 2014), Experiment 2A utilized a sorting task as the final memory test. Participants were given all of the studied items and asked to sort them into the appropriate list. The critical measure was performance on List 3 (which was insulated from primacy and recency effects, along with having an equal number of retrieval and distractor tasks prior to the final test). This design allowed for insight into the underlying mechanisms based on category confusion (e.g., putting a List 2 word in the List 3 category) and overall memory performance (e.g., proportion of correct responses for List 3) without the confound of a different number of retrieval events between the target list (List 3) and the final test (as is the case in designs where participants perform either *all* retrieval or *all* distractor tasks). Experiment 2B is a follow-up on Experiment 2A, aimed at eliminating some of the interdependencies inherent in the dependent measures of the Experiment 2A design. Experiment 2B is an abbreviated, 3-list version of Experiment 2A in which participants are only tested on the middle list.

[10] Prior work (e.g., Divis & Benjamin, 2014) has consistently used a between-subjects design, which is costly in terms of resources and lacks a complete measure of the manipulation within a given participant.

Experiment 2A

Method

Participants

Thirty-three undergraduates from the University of Illinois at Urbana-Champaign participated in this experiment for course credit. The data from three participants were dropped (they failed to follow the semantic retrieval instructions), leaving a total of 30 participants' data for analysis.

Design

Type of intervening task (semantic retrieval or counting distractor) was manipulated within subjects, and task order was counterbalanced across subjects. Memory performance was measured via a sorting task. Participants were given all of the studied words (along with new, unstudied words) and were asked to sort them into their appropriate category (e.g., studied in List 1, unstudied, etc.).

Materials

Fifty words (average word length = 5.08 letters, *SD* = 1.40 letters) from the University of South Florida Free Association Norms (Nelson et al., 1998) were used to create the study lists, with 10 words randomly assigned to each list. None of the studied words were related to the categories used for the semantic retrieval task; they were also unrelated to each other. An additional 10 words were used to create an "unstudied" category.

Procedure

This experiment was completed in the same environment and with the same software as Experiment 1. Before the study session, participants were presented with instructions informing them that they would be studying lists of words for a later memory test and that it was important to remember the words and to which list they belonged.

During the study session, participants studied five lists of words (labeled numerically), separated by an intervening task. Words were presented one at a time for 4 seconds each, with an interstimulus interval of 500 ms. The intervening task was either a distractor counting task or a semantic retrieval task (similar to those used in Experiment 1). In the distractor task, subjects counted aloud backward by 3s from a randomly generated 3-digit number for 30 seconds. This was repeated two more times with a new randomly generated 3-digit number each time for a total of 90 seconds (three 30-second blocks). In the semantic retrieval task, participants first completed one 30-second counting block (just as in the distractor task) and then generated exemplars from a given semantic category (vegetables or 4-legged animals) for 60 seconds. Responses were typed into boxes on the computer screen and participants could not change a response once it was entered. Participants completed two retrieval tasks across the experiment; which category occurred first (vegetables or 4-legged animals) was randomly determined. While each participant had both distractor and semantic retrieval intervening tasks, the order was counterbalanced across subjects, creating two orders (DRDR and RDRD)[11], as shown in Figure 5. After finishing studying List 5, participants spent 60 seconds completing a series of simple addition problems (adding two 2-digit numbers) prior to beginning the test phase.

[11] Having both retrieval and distractor tasks for each subject allows for a within-subject measure of the manipulation. Importantly, the critical measure will focus on performance on List 3 (which is insulated from primacy and recency). Are there fewer intrusions onto List 3 from nearby lists (i.e., List 2 or List 4) when they are separated by a retrieval as opposed to distractor task?

In the test phase, participants completed a memory sorting task on paper. Participants were given all 50 words they studied in the study phase along with 10 new words from the "unstudied" category (all arranged alphabetically). Their task was to assign each word to its proper category (List X or Unstudied). Participants were instructed to use all 60 words and fill in ten words in each category (without repeating a word). Participants completed this test by filling in their responses by hand; no time limit was imposed.

Results and discussion

A priori analyses were planned to (1) examine the (expected) list serial position effect on overall accuracy and (2) examine interlist intrusions[12] onto List 3. List 3 was intentionally selected for analysis because it should be most insulated from confounds due to primacy and recency effects (along with having an equal number of retrieval and distractor events between it and the test, regardless of condition). Specifically, I was interested in whether more near-list intrusions onto List 3 occurred when the lists were separated by a distractor relative to a retrieval task, regardless of whether those intrusions originated from List 2 or List 4. I also planned to examine the simple effects of intervening task for both near-lists, since we found proactive interference (i.e., intrusions from List 2) but not retroactive interference (i.e., intrusions from List 4) in previous work (Divis & Benjamin, 2014; Experiment 1 in the current manuscript). The measures of near-list intrusions were interdependent with accuracy and other intrusion measures due to the nature of the sorting task—that is, number of possible intrusions was limited by

[12] While the intrusions that are induced by a classic free recall paradigm (e.g., as in Experiment 1 in the current manuscript and Divis & Benjamin, 2014) may have different properties than intrusions induced in the sorting task used here, I will continue to use the term "intrusions" rather than a new term such as "confusions." Most importantly, they both are an indication of the same phenomenon—less distinct boundaries between lists—regardless of paradigm. Participants are committing errors indicating they are less able to differentiate which items belonged to which list. However, the design of Experiments 2A and 2B lead to a higher rate of intrusions, making it easier to test the predicted reduction in intrusions (interlist intrusions tend to be near floor in free recall tests).

accuracy performance on the target list and intrusion source list. Follow-up analyses were conducted to try to account for that discrepancy; Experiment 2B directly addresses it. Exploratory analyses were also conducted to examine overall near-list intrusions and List 1 accuracy performance. Significance levels for all statistical tests were set at an $\alpha < .05$ level. Random effects models were fit as in Experiment 1. See Appendix A for model fitting details; see Appendix B for Bayes factors.

Accuracy

Overall accuracy by condition is shown in Figure 6a. The expected U-shaped serial position curve showing primacy and recency benefits was found. Both the DRDR counterbalancing condition ($R^2 = .873$) and the RDRD counterbalancing condition ($R^2 = .780$) were well-fit by a quadratic regression line (see Figure 6b).

Near-list intrusions on List 3

Intrusion analyses were planned *a priori* to focus on the middle list so they would not be contaminated by the anticipated primacy and recency effects. Near-list intrusions for List 3 (either List 2 items intruding on memory for List 3 or List 4 items intruding on memory for List 3) are shown in Figure 7a. The base model included task (retrieval or distractor) and source list (List 2 or List 4) with random intercepts for subjects to predict near-list intrusions on List 3. The model revealed a main effect of task, with more intrusions occurring when the lists were separated by a distractor task compared to a retrieval task ($t(30) = 3.006$, $p = .005$). No significant simple effect of task for List 2 was found ($p = .167$); however intrusions from List 4 were more likely when preceded by a distractor relative to retrieval task (*t*(59.96) = 2.800, *p* = .007).

Notably, near-list intrusions might be influenced by additional factors such as proportion correct on the target list or nearby lists. For example, a case where only two items were correctly identified as belonging to List 2 has more List 2 items available to intrude on other lists than a case where nine items were correctly identified as belonging to List 2. For that reason, models were built off of the base model described above to see if adding additional effects would significantly improve the fit of the model (see Appendix A for details). The best fit model accounted for performance on the intrusions' source list in addition to the fixed effects of task (retrieval or distractor) and near-list (List 2 or List 4) and random effect of subject to predict intrusions onto List 3. Once again, the model revealed a main effect of task, with more intrusions occurring between lists divided by a distractor relative to a retrieval task (*t*(29.28) = 2.941, *p* = .006). A simple effect of task for both List 2 (*t*(47.95) = 2.068, *p* =.044) and List 4 (*t*(49.64) = 2.195, *p* = .033) revealed more intrusions occurring in the distractor compared to retrieval condition. When compared with the base model, these results indicate that not including the influence of performance on the intrusion source list was exaggerating the simple effect of task in List 4 but suppressing the simple effect of task in List 2.

Exploratory analyses

The main goal before running this experiment was to look at the pattern of interference in the middle list, based on intervening task. However, the full dataset also allows for exploration of other relevant questions. Figure 7b shows near-list intrusions for all five lists. All things being equal (e.g., no influence from serial position effects), more intrusions should occur when lists are separated by a distractor (as opposed to retrieval) task. Numerically, this occurs in all cases except List 1 intrusions onto List 2. A model predicting near-list intrusions (for any of the 5 lists) from the fixed effect of intervening task (retrieval or distractor) and random intercepts for subjects showed that significantly fewer intrusions occur when the lists are separated by a retrieval as opposed to distractor task ($t(240) = 2.602$, $p = .010$). Figure 7c summarizes the proportion correct and near-list intrusion data, collapsing across list and condition.

List 1 performance when followed by a retrieval (as opposed to a distractor) task is of interest when examining the predictions of retrieval-induced distinction compared to inhibition or interruption of consolidation hypotheses. List 1 is the only list that (1) cleanly leads to different predictions (as outlined below) and (2) is insulated from proactive interference.

The two components driving the effect of retrieval on memory performance for earlier material from a retrieval-induced distinction perspective are (1) list segregation leading to reduced interlist interference and (2) increased mental distance between the studied list and the criterion test (more retrieval events between study and test will lead to more spreading out of the representations; i.e., greater contextual disparity). Because an equal amount of retrieval and distractor task instances occurred between List 1 and the test in Experiment 2A, the latter point should not be a factor.

While the nature of the sorting task will inherently lead to a more jumbled test context (as earlier list contexts are reinstated), if anything retrieval-induced distinction via context change weakly predicts *increased* List 1 performance in the RDRD compared to DRDR counterbalancing conditions (due to a reduction in retroactive interference). In contrast, an inhibition or interruption of consolidation perspective predicts that retrieval serves to reduce memory performance for items prior to the retrieval task (because those items are inhibited or not properly consolidated); it would strongly predict *reduced* performance on List 1.[13] That expected reduction in List 1 performance is not found. The best-fit model predicting accuracy on List 1 included the fixed effect of task following List 1 (retrieval or distractor) and random intercepts for subjects. It revealed no reliable difference in List 1 accuracy based on subsequent intervening task (*z* = .082, *p* = .934). Interpretation of this null finding should be taken with caution, but it hints toward a failed strong prediction from an inhibition or interruption of consolidation perspective and a failed weak prediction from a retrieval-induced distinction perspective[14]. Once again, this experiment was designed to focus on *intrusions*—not List 1 performance and interpretation of the results should be taken with caution.

Summary

As expected, near-list intrusions were reduced when lists were separated by a semantic retrieval task compared to a distractor counting task. These results support the hypothesis that interleaved retrieval serves to make lists more distinct from one another and thus less likely to interfere. Notably, this design utilized a weaker manipulation[15] of the interleaved task than prior work and still found results consistent with retrieval-induced distinction. Instead of completing the same intervening task across the entire study session (a continuous buildup of four tasks), participants alternated tasks within the study session (a continuous buildup of only one task). The

[13] Due to the nature of the counterbalancing conditions and alternating the retrieval and distractor tasks, when List 1 is followed by a retrieval task, List 2 is then followed by a distractor task (and vice versa). From an interruption of consolidation or inhibition perspective, when List 1 is followed by a retrieval task, memory for List 1 is weakened. But since List 2 is then followed by a distractor task, memory for List 2 should *not* be similarly weakened. Therefore a reduction in retroactive interference (List 2 interfering with List 1) is not expected here (unlike in the retrieval-induced distinction explanation).

[14] Although the prediction of reduced retroactive interference was supported.

[15] A stronger manipulation (e.g., RRDD or DDRR) might lead to stronger effects; but from a retrieval-induced distinction perspective, it would also introduce the additional confound of different relative placement within the contextual stream (i.e., mental distance) between the list of interest (List 3) and the criterion test (those in the DDRR condition would have a greater dissimilarity in List 3 and test contexts than those in the RRDD condition).

effects of retrieval seen in free recall paradigms generalize well to intrusions in list sorting, supporting predictions based on retrieval-induced distinction.

Experiment 2B

Experiment 2A showed that near-list intrusions are more likely to occur when the lists are separated by a distractor compared to retrieval task. However, the design of Experiment 2A inherently led to interdependence among the memory measures. In an effort to replicate the intrusion results and experimentally (rather than just statistically) address that potential confound, I collected preliminary data on an experiment identical to that of Experiment 2A, except that participants were only tested on List 3 items. If participants are not required to sort every single word into the appropriate category, they will not be constrained by what is “left.” Unfortunately, this design failed to elicit the expected response. When not forced to categorize all the words, participants appeared to rely on overall *strength* of the word rather than *list category membership*. Only 6 participants (24%) successfully assigned more words from List 3 than from any single other list (75% inaccurately assigned more words from at least one single list other than List 3; 1% assigned equal amounts), indicating they were having difficulty successfully falling into List 3 words and not relying on overall memory strength. A similar pilot study utilizing a forced-recognition test rather than free recall revealed the same inability to successfully access List 3 items; performance was at or below chance.

These failures motivated a reconsideration of the number of study-task cycles. In an attempt to make access to the target list more successful, Experiment 2B compromised the benefits of a five-list design[16]. Participants studied *three* lists of words, were given all of the studied words, and were asked to pick out the words belonging to the middle list.

Method

Participants

Fifty undergraduates from the University of Illinois at Urbana-Champaign participated in this experiment for course credit. The data from 10 participants were dropped (they either failed to follow the semantic retrieval instructions or were familiar with the manipulation prior to beginning the experiment), leaving a total of 40 participants' data for analysis.

Design

Type of intervening task (semantic retrieval or counting distractor) was manipulated within subjects, the order of which was counterbalanced across subjects. Memory performance was measured via a sorting task.

Materials

Forty-five words (average word length = 5.56 letters, *SD* = 1.15 letters) from the University of South Florida Free Association Norms (Nelson et al., 1998) were used to create the study lists. Fifteen words were randomly assigned to each list. None of the studied words were related to the semantic retrieval category; they were also unrelated to each other.

[16] That is, insulation from primacy and recency effects and an equal number of retrieval and distractor tasks between the target list and criterion test.

Procedure

The procedure was identical to that of Experiment 2A, except where noted. Participants studied three lists of words, separated by an intervening task. The intervening task was either the distractor counting task or semantic retrieval task used in Experiment 2A. The semantic retrieval category was "4-legged animals." Order of the intervening task was counterbalanced across participants. Participants in the RD condition studied List 1, completed the retrieval task, studied List 2, completed the distractor task, and then studied List 3; those in the DR condition completed the distractor task after List 1 and the retrieval task after List 2. After finishing studying List 3, all participants spent 90 seconds completing simple addition problems.

In the test phase, participants were given all 45 studied words (in alphabetical order) and asked to identify the 15 words belonging to List 2. Participants were asked to provide 15 words, with no repetitions. Responses were completed on paper and no time limit was imposed.

Results and discussion

Analyses were planned to look at the overall main effect of task, along with the simple effects of task for each source list (List 1 and List 3) for intrusions onto List 2. Figure 8a shows proportion of intrusions onto List 2 by source list and intervening task. Because List 2 and the criterion test were separated by *either* an intervening retrieval or distractor task, performance correct on List 2 should vary based on condition (DR or RD); therefore proportion correct on List 2 was also assessed. See Figure 8b for accuracy performance on List 2 by condition. See Appendix A for details on model fitting; see Appendix B for Bayes factors.

Interlist intrusions onto List 2

The best fit model predicting interlist intrusions onto List 2 included fixed effects for intervening task (retrieval or distractor) and source list (List 1 or List 3), along with random intercepts for subjects. The model revealed a significant main effect of task, with overall reduced intrusions for the retrieval task relative to the distractor task (t(80) = 2.017, *p* = .047). The model also revealed a simple effect of task for intrusions from List 1, with more intrusions occurring when the lists were separated by a distractor compared to retrieval task (*t*(80) = 2.131, *p* = .023). While numerically fewer intrusions from List 3 onto List 2 occurred when separated by a retrieval relative to distractor task, that simple effect was not reliable.

As expected and Experiment 2A showed, primacy and recency effects can muddle measurements of near-list intrusions for information at the beginning or end of the study session. To avoid the interdependence confound of Experiment 2A, Experiment 2B did not collect performance measures for Lists 1 or 3, so a direct measure of primacy and recency is not possible here. The overall expected pattern of results (reduced intrusions when the lists were separated by a retrieval task) appeared. This pattern was significant for intrusions from List 1; while numerically (but not statistically significantly) still apparent for intrusions from List 3 (possibly the effect was washed out due to the noise introduced by recency effects).

Accuracy

The best fit model predicting accuracy included the fixed effect of condition (DR or RD) and random intercepts for subject. Numerically, performance was higher in the RD compared to DR condition, but that effect was also not reliable (*z* = 1.351, *p* = .177).

Failure to find a statistically significant difference (but the appropriate numerical pattern) was not surprising. Early work in this line of research indicated that the magnitude of the effect of interleaved retrieval on prior memory increases with more lists and more retrieval events. Early pilot work with fewer lists also showed numerically correct but unreliable patterns of results.

Summary

Taken together, Experiments 2A and 2B indicate that, overall, interlist intrusions are reduced when lists are separated by a retrieval event compared to a distractor event. Experiment 2A controlled for primacy and recency effects along with mental distance between the target list and criterion list (an equal number of retrieval and distractor tasks occurred between the target list and criterion test) at the cost of interdependence among the memory measures. Experiment 2B controlled for interdependence among the memory measures at the cost of influence from serial position effects and a different number of retrieval and distractor tasks between the target list and criterion test based on condition. While neither experiment is perfect, they provide converging evidence that interleaved retrieval leads to list segregation and a reduction in interlist intrusions.

Experiment 3: Interference and Facilitation with Complex Material

Experiment 3 was designed to simultaneously examine the effects of interleaved retrieval on both interference and facilitation within one design while using more complex, narrative text materials on the same topic.

Influence of interleaved retrieval on related material

Interleaved retrieval appears to spread out memory representations, making them more distinct and less likely to interfere with each other. By this retrieval-induced distinction interpretation, it should also be more difficult to *beneficially* connect information between those memory representations. Interleaved retrieval should lead to a reduction in interlist facilitation. When two semantic associates are seen across a study session, seeing the second word “reminds” one of seeing the first word and leads to better memory for that first word (the *reminding effect*; Tullis, Benjamin, and Ross, 2014). For example, studying “king” and later studying “queen” leads to better memory for “king” (compared to when “queen” was not studied). A retrieval event between the two related words should make it harder to achieve the beneficial connection between those two associates and thus reduce the reminding effect.

In two pilot studies, I started to test this prediction but was unable to find a design that elicited strong effects. The three primary challenges are: (1) This is a prediction of a difference of differences (i.e., a reduction in the reminding effect), which requires a high degree of power. (2) It is easiest to study the effects of interleaved retrieval when they build across multiple studied lists. However, one can generally only have related words across two lists. (3) Because the reminding effect is usually small, it is possible to run into floor effects as the reminding effect is reduced.

In the first pilot study, 58 participants studied two lists of words, separated by a distractor counting task or semantic retrieval task. The lists consisted of 5 related words and 5 unrelated words. The words were related across lists (e.g., “king” in List 1 was related to “queen” in List 2). Participants were then tested on List 1. Figure 9a shows proportion correct on List 1 as a function of intervening task and material type. The difference between performance on related and unrelated words is the reminding effect. Unfortunately, while there was a numerical drop in the magnitude of the reminding effect in the control condition (reminding effect = .09) relative to the retrieval condition (reminding effect = .05), it was not significant. This design demonstrated the predicted drop in the reminding effect, but it was too weak to be reliable.

In an effort to overcome the weaknesses of the first pilot experiment, the second pilot experiment switched to an abbreviated list-before-last paradigm. One hundred ten participants studied two lists of words, were either tested on List 1 or completed a distractor counting task, studied a third list of words, and then were tested on List 2. Once again, half the words were related across List 2 and List 3 and half were unrelated. Figure 9b shows proportion correct on List 2 as a function of intervening task and material type. While there was a significant reminding effect for the control condition and no reminding effect for the retrieval condition, the difference between the reminding effects was not significant. Once again, the pattern of results was in line with predictions but the effect of interest was too weak to detect (a difference of differences).

Across two experiments with a combined total of 168 participants, I found a consistent reduction in the reminding effect when lists were separated by a retrieval relative to distractor task. However, neither experimental design elicited a strong, significant effect. In Divis & Benjamin (2014), we found evidence indicating that the effects of retrieval on overall accuracy performance appeared stronger with complex narrative materials than simple lists of words, suggesting that using more complex materials might also lead to a stronger effect here.

Experiment 3 motivation

This experiment ambitiously attempted to combine effects from multiple previous studies into a single large experiment using more complex narrative text materials (eye witness testimonies of the same crime). The materials were designed to include information that would be likely to cause both interference and facilitation across texts. Furthermore, the tests were designed to assess both overall accuracy performance and confusions between the texts. I predicted that interleaved retrieval would lead to both a detrimental reduction in the benefits of having similar information repeated across the texts and a beneficial reduction in the consequences of having contradictory information across the texts.

The greater complexity in this design compromised much of the experimental control found in the earlier experiments. Having multiple types of information (e.g., contradictory) tied across texts could interact in unexpected ways. The longer, multiple memory tests after the study session could also add noise to our measures[17]. However, how the effects of interleaved retrieval transfer to more complex, "real world" settings such as classrooms is an intriguing question and the natural next step in seeing whether this basic memory research transfers to more applied settings.

[17] In fact, Divis & Benjamin (2014) showed that a significant delay between the study session and criterion test eliminated the effects of interleaved retrieval.

Method

Preregistration

This experiment was preregistered. The preregistration documents outline the basic predictions, expected confounds and challenges with the design, the materials and code, number of subjects to be run, stopping rules, and rules for excluding data. (They can be found at https://osf.io/dvy67/?view_only=29b737600e0a4ba9abb61715b384e00c).

Participants

One hundred ninety-nine undergraduates at the University of Illinois at Urbana-Champaign completed this experiment for course credit. The data from 13 participants were dropped due to computer error or for failing to follow the semantic retrieval instructions, leaving a total of 186 participants' data for analysis.

Design

Type of task (semantic retrieval or distractor) and eye witness study position tested during the multiple choice test (first or last) was manipulated between subjects, for a total of four conditions. Order of the eye witness testimonies (Tony-Gary-Chris-Steven or Steven-Chris-Gary-Tony) was also counterbalanced between subjects. Each eye witness testimony contained information that was unique to that eye witness testimony and information that was tied to details from another eye witness testimony (contradictory, repeated, or superadditive). Memory was assessed via two consecutive tests: (1) a multiple choice test on either the first or last eye witness testimony and (2) a short answer test on all the eye witness testimonies. See Figure 10 for a diagram of the experimental design.

Materials

Four eye witness testimonies (Tony, Gary, Chris, and Steve) about a minor crime were created for this experiment (average number of words = 360, *SD* = 41 words; 3 paragraphs per testimony). The contradictory, repeated, and superadditive information was always paired between either Tony and Gary or Chris and Steve. See Appendix C for the exact testimonies.

The multiple choice test consisted of a total of 17 questions (4 questions on contradictory information, 3 questions on superadditive information, 6 questions on repeated information, and 4 questions on unique information). Since the eye witness testimony from Tony or Steve was always in the first or last study position, only their testimonies were queried during the multiple choice test. See Appendix C for all the multiple choice questions used.

The short answer test had two parts: a matrix section and a unique information section. The matrix section had a total of 26 questions. For each question, participants were instructed to first answer the question based on *all* of the eye witness accounts and then say *which individual* eye witness provided each piece of information. The questions were always on information that was contradictory, repeated, or superadditive. See Appendix C for the exact instructions used and questions asked. The unique information section of the short answer test consisted of a total of 12 (3 for each eye witness) free response questions on information that was unique to an eye witness. See Appendix C for the questions asked. Notably, all of the information queried in the multiple choice test was also queried in the short answer test.

Procedure

Once again this experiment was completed in the same environment and with the same software as in Experiments 1, 2A, and 2B. Prior to starting the experiment, participants were presented with the following instructions:

> *Thank you for participating in this study. In this experiment, we would like you to imagine that you are a detective collecting eye witness testimonies of a crime. You will read through four eye witness testimonies of the same crime. Each eye witness had a different perspective on the crime and was able to see different aspects of the crime.*
>
> *Your job is to (1) Remember the* ***details*** *of each eye witness testimony and* ***who*** *(which eye witness) reported those details. It's important you remember both the details of the crime and any other information the eye witness provides (whether or not it seems important to the crime). (2) Build an* ***overall*** *picture of what happened during the crime. At the end of the experiment, your memory for the details about each eye witness and the crime will be tested.*
>
> *During the experiment, you will NOT be allowed to go back and reread a prior screen after advancing to the next screen. Therefore, it is important that you take your time and read through the eye witness accounts carefully. The eye witness testimonies will be broken up with short tasks (e.g., counting backward).*

During the study session, participants were informed of the name of the eye witness before reading his testimony. Eye witness testimonies were presented one paragraph at a time, and participants were allowed to read at their own pace (the only criterion was that they spend at least 15 seconds on each paragraph). Participants were not allowed to go back and reread a testimony after advancing to a new screen. Between studying each eye witness testimony, participants were asked to complete an intervening task. The intervening tasks were the same as those used in Experiments 2A and 2B: either three sets of the 30-second distractor counting task or the 30-second distractor task followed by the 60-second semantic retrieval task (generating examples of vegetables, four-legged animals, or sports). After finishing the last eye witness testimony, participants solved simple 2-digit addition problems for 90 seconds.

The multiple choice test was completed on the computer. The order of the multiple choice questions was randomized and the answer options were scrambled. Participants selected an answer by clicking on it; the answer was highlighted and the experiment advanced. Participants were given as much time as they needed to complete the multiple choice test; they were not allowed to go back and change their answers.

After completing the multiple choice test, participants were given the short answer test, which was completed via paper and pen. Once again, no time limit was imposed. After finishing the short answer test booklet, participants were debriefed and given course credit for their participation.

Results

As outlined in the preregistration documentation, I expected to find that eye witness testimonies separated by retrieval tasks exhibited evidence of increased distinction[18]. I predicted less interference but also less facilitation between segregated testimonies. In line with Divis and Benjamin (2014), I also expected overall accuracy to be poorer for the first eye witness and better for the last eye witness in the retrieval relative to distractor conditions. However, the interaction between the reduction in interference and facilitation and the accuracy as a function of the testimony serial position cannot be perfectly controlled for in this design, and I anticipated

[18] In line with preregistration plans, these analyses were also run after dropping the data from the first 10 pilot subjects. Similar results were found; see Table 1.

that it may introduce an additional sources of noise into the data. For example, subjects in the retrieval condition are expected to have overall poorer memory for the first eye witness but they are also expected to have less interference from contradictory information in the second eye witness testimony (which would boost memory performance for the first eye witness).

As also addressed in the preregistration documentation, the multiple choice test is the cleanest measure, especially for overall accuracy. The short answer test is a weaker measure because it comes after the multiple choice test (an additional retrieval event and a brief delay) and information tested during the multiple choice test is also tested during the short answer test. Therefore the results of the short answer test must be interpreted with considerable caution, and are probably more aptly treated as exploratory in nature.

Once again, significance levels for all statistical tests were set at an $\alpha < .05$ level. Except where noted, statistical analyses were run using mixed effects models as in Experiments 1, 2A, and 2B. See Appendix A for information on model fitting; see Appendix B for Bayes factors.

Analyses of the individual matrix section of the short answer matrix test were run on data from only 173 subjects (data from 13 subjects was unusable due to failure to follow the instructions on how to fill out the matrix). One rater scored all of the short answer test booklets. An additional, trained rater independently scored a random subset (16) of the test booklets, with the criteria that each booklet had to have complete data and that two booklets were selected from each of the 8 conditions (including the counterbalancing manipulation). Inter-rater reliability scores were high (r = .973 for the matrix portion; r = .964 for the unique short answer section).

Accuracy (Multiple Choice Test)

The best fit model for proportion correct in the multiple choice test included fixed effects for intervening task (retrieval or distractor) and eye witness testimony position tested (first or last), along with random intercepts for subject and question. See Figure 11.

All types of information. The model revealed no significant differences in proportion correct between the retrieval and distractor conditions for either the simple effect of the first position or the simple effect of the last position when collapsed across all types of information. Overall accuracy was relatively high (proportion correct = .861).

Unique information. The model also revealed no significant differences in proportion correct between the retrieval and distractor conditions for either the simple effect of the first position or the simple effect of the last position when selecting only the questions querying unique information. Accuracy was near ceiling for these questions (proportion correct = .939).

Repeated information. No significant differences in proportion correct between the retrieval and distractor conditions were revealed by the model for either the simple effect of the first position or the simple effect of the last position when selecting only the questions on repeated information. Accuracy was also near ceiling for the questions about repeated information (proportion correct = .923).

Contradictory information. When analyzing only questions on contradictory information, performance was significantly lower in the distractor relative to retrieval condition for the simple effect of the first position ($z = 2.253$, $p = .024$); the model revealed no significant differences for the simple effect of the last position. Accuracy was lower than for other types of information (proportion correct = .742).[19]

Discriminability (Short Answer Test: Individual Matrix Section)

[19] All superadditive information is addressed in Appendix D.

Discriminability was analyzed via d’ scores for the matrix section of the short answer test where participants indicated which individual eye witness said each piece of information. A response counted as a “hit” if a participant indicated that an eye witness talked about that question when the eye witness actually did mention information pertinent to that question. A response counted as a “false alarm” if the participant indicated that an eye witness mentioned that question when he actually did not. This analysis only factored in whether the participant indicated that the eye witness talked about that information (not whether the information given was accurate). In order to be able to calculate d’ scores, perfect hit rates were rescored to the halfway point between a perfect score and missing one item (.99); the same was done for perfect false alarm rates (.01; Snodgrass & Corwin, 1988). d’ scores were calculated collapsing across all the conditions for each subject. A t-test[20] revealed no reliable difference between d’ scores in the retrieval and distractor conditions. Overall, discriminability scores were high; see Figure 12.

Foil Endorsement (Multiple Choice Test)

When questions referred to contradictory information paired across two eye witness testimonies, the multiple choice response options included the correct answer for the eye witness tested along with a foil answer (what the other eye witness said). These analyses examine endorsement rates for those foils. See Figure 13.

The best fit model for proportion foils endorsed on contradictory information in the multiple choice test included the fixed effects of intervening task (retrieval or distractor) and eye witness test position (first or last), along with random intercepts for question. Overall, the model revealed a marginal effect of intervening task, with fewer foils endorsed in the retrieval relative

[20] Mixed effects models would require too many adjustments for perfect hit rates and perfect false alarm rates to be a reasonable statistical analysis choice.

to distractor condition (*z* = 1.857, *p* = .063). The same pattern held for the simple effect of the first eye witness position (*z* = 1.861, *p* = .062) but not the last eye witness position.

Misattributions (Short Answer Test: Individual Matrix Section)

These analyses look at misattributions in the matrix section of the short answer test where participants indicated which individual eye witness said each piece of information. A misattribution occurred when a participant attributed information to the wrong eye witness. See Figure 14.

All data. The best fit model using all of the data included a fixed effect for intervening task (retrieval or distractor) and random intercepts for question. It revealed an overall effect of fewer misattributions in the retrieval relative to distractor condition (*z* = 2.318, *p* = .021). The same pattern was found when just looking at contradictory information (*z* = 1.994, *p* = .046) and, marginally, when only examining repeated information (*z* = 1.804, *p* = .071).

Only untested pairs. Before taking the short answer test, participants were tested on one of the eye witnesses in the multiple choice test. Information in the testimonies was connected between pairs of eye witnesses. Therefore half of the information in the matrix section of the short answer test had already been queried prior to the participant seeing it again in the short answer test. One way to attempt to account for that potential confound is to remove all previously tested information from the analysis. After doing so, the model that best fit the data included a fixed effect for intervening task (retrieval or distractor) and random intercepts for subject and question. Overall, significantly fewer misattributions occurred in the retrieval relative to distractor conditions (*z* = 2.499, *p* = .012). The pattern held for repeated information (*z* = 2.289, *p* = .022) but not contradictory information.

Supplementary Analyses

The complex experimental design of Experiment 3 led to tests of strong and weak theoretical predictions. Analyses of the weakest tests are briefly described below and can be found in Appendix D.

The superadditive questions were originally intended to provide another way to look at the effect of interleaved retrieval on facilitation. If one eye witness said the suspect fled north and another said he went down Washington Ave., the full answer (north on Washington Ave.) can only be achieved by combining the two pieces of information. Perhaps reading that the suspect went down Washington Ave. would remind one that the earlier eye witness had said he went north, leading to memory facilitation. However, the superadditive information could also backfire and provide interference. Which eye witness said “north” and which said “Washington Ave.”? Could those actually be interpreted as *different* things instead of two components of the same correct response? Furthermore, there was some question as to what actually makes something “superadditive” (refer to Appendix C for the details used in this experiment). The analyses of all superadditive information can be found in Appendix D.

One consequence of the design of this experiment was that the short answer test was a weaker measure due to the fact that it followed the multiple choice test (which also tested some of the same information). Analysis of memory accuracy in the short answer test revealed no results of consequence. See Appendix D.

Discussion

One of the intentions when designing this experiment was to combine and push effects found with more basic material onto more complex, ecologically valid material. Experiments 2A and 2B (see also Divis & Benjamin, 2014) showed reduced interference between word lists separated by a retrieval event. Pilot studies using related words showed a trend toward reduced facilitation between word lists separated by a retrieval event, but the interaction was never significant. We found strong overall accuracy effects from interleaved retrieval with narrative texts for the reading comprehension section of the ACT (Divis & Benjamin, 2014). Would more complex material that had contradictory and repeated information built in across the texts also be susceptible to effects of interleaved retrieval? These results suggest that may not be the case.

Why did the results of this experiment not pan out as expected? The greater complexity of the experiment could have led to a large enough increase in inter-individual or inter-item variance that it washed out the effects of interleaved retrieval. Perhaps having directly contradictory and repeated information in addition to the unique information interacted with the normal effects of retrieval-induced distinction. Another possibility is that performance was too high in this experimental design. Possible ceiling effects were particularly concerning for accuracy performance on the multiple choice test. Perhaps interleaved retrieval has the largest effect when learners do not already know the studied material well. All of these differences are possible contributors to the predominantly null effects seen in this experiment, but these limitations may not tell the whole story.

One of the biggest differences between this design and the Divis and Benjamin (2014) experiment using ACT texts is that the materials are strongly connected. All of the texts are about the same crime, just from different perspectives. Many of the elements and visual imagery strongly overlap. Every time a participant read an eye witness testimony, it was about the same crime and had similar elements to previous testimonies. That is not to say that some of the information was not distinctive and that having interleaved retrieval could not theoretically have enhanced that distinction—but rather that the powerful shared imagery might have overwhelmed many of the effects of retrieval-induced distinction. This is a quality that is likely to be apparent in *any* complex design trying to achieve the goals of this experiment. In order to have repeated or directly contradictory elements, those elements must overlap across texts.

Another potential factor is how concrete or abstract the information being studied is. In this experiment, all the elements described were concrete and easily imagined (i.e., crime scene). Would the effects have been similar with more abstract materials (e.g., an essay on the concept of justice)? If the vivid, shared imagery was one of the primary reasons many of the effects of retrieval-induced distinction were washed out in this experiment, using such concrete material may have aggravated that problem. Perhaps more realistic, educationally-relevant material that focuses on abstract concepts would not be as easily influenced by shared imagery. This prediction is particularly important for the expected reduction in facilitation and interference between material separated by retrieval (overall accuracy should be better for concrete relative to abstract information).

Is there a larger distinction effect for abstract relative to concrete material? The only study to date that can begin to address this question is Experiment 2A in the current manuscript. Importantly, the relative concreteness of words was not counterbalanced in that design and there were relatively few subjects and items. However, as a first pass to test this prediction, I ran analyses on the concreteness of the words. Concreteness scores were assigned to each of the words using the norms outlined in Brysbaert, Warriner, and Kuperman (2014). The scale ranged from 1 (abstract) to 5 (concrete). Words that had an average score of 3 or below were binned into the “abstract” category; words higher than 3 were categorized as “concrete”. Twenty-five percent of the words used in Experiment 2A were abstract by this measure; 28% of the near-list intrusions were from abstract words, indicating that the abstract words were, at most, slightly more confusable than the concrete words. Thirty percent of the near-list intrusions in the retrieval condition were from abstract words, compared to 26% in the distractor condition. However, this measure is confounded by the overall *fewer* number of near-list intrusions in the retrieval relative to distractor condition (perhaps the most challenging words were the ones that still intruded in the retrieval condition). This exploratory analysis on the relative concreteness of intrusions from Experiment 2A does not support the idea that the effects of retrieval-induced distinction are stronger for abstract information (likewise, it also cannot provide definitive evidence that relative concreteness has *no* effect on retrieval-induced distinction).

Notably, the only portion of the principal measure of memory performance (the multiple choice test) that revealed a difference between intervening retrieval and control tasks was for contradictory information said by the first eye witness[21]. Participants in the retrieval condition demonstrated better memory for the first eye witness and were less likely to incorrectly endorse the foil from the second eye witness when answering questions about contradictory information between the first two eye witnesses. The same pattern did not hold for the last eye witness testimony. Perhaps retrieval better insulates from misinformation that comes after the retrieval event rather than prior to it (i.e., it reduces the likelihood of information from the second eye witness being confused for information from the first eye witness but does not similarly protect the likelihood of information from the third eye witness being confused for information from the

[21] Although it should be highlighted that memory performance for the other types of information was at ceiling, so lack of an effect is not particularly compelling evidence for either the null or alternative hypotheses.

fourth eye witness).[22] Alternatively, the last eye witness may have benefited from a recency effect, regardless of condition.

Summary

This experiment highlights some important boundaries in the effect of interleaved retrieval. In this setting, the *overall accuracy* effects of retrieval-induced distinction disappeared. There appeared to be fewer *confusions* between the stories, especially for contradictory information, but the effect was small. This is the first time interleaved retrieval has led to such a convincingly null effect on overall memory accuracy. The biggest contenders for what caused that lack of effect are (1) interleaved retrieval is not effective on material that is already very well learned and (2) interleaved retrieval is not effective on material that has many overlapping elements and shared context. While those two factors also may have dampened the effect of retrieval-induced distinction on confusions between memories, it did not completely wipe it out, suggesting that even in highly similar, well-learned settings, interleaved retrieval can still be beneficial for reducing source memory errors.

One open question has been how well retrieval-induced distinction effects apply to educational settings. In Divis and Benjamin (2014), we found strong accuracy effects with unrelated ACT reading comprehension texts. Here, I found strong null effects for accuracy with well-learned stories about the same event. However, there was still a trend for reduced confusions between the stories. Perhaps interleaved retrieval is only beneficial for overall memory accuracy in settings where the material is challenging and about different topics. During study hall, students might benefit by breaking up studying for their math test and chemistry test

[22] Notably, this pattern of effects is the opposite of what an inhibition or interruption of consolidation perspective would predict.

with a retrieval event, but not between studying two chapters in their history book (though that still might help them keep them from confusing events in the second chapter for events in the first chapter). Future experiments could directly address these questions by (1) using more difficult material that is likely to be less well-learned by the time of the test (or by imposing a shorter time limit) or (2) manipulating how many overlapping elements the materials include (materials less strongly connected may show stronger retrieval-induced distinction effects).

General Discussion

Interleaved retrieval consistently led to overall reduced performance on pre-retrieval material and enhanced performance on post-retrieval material in prior work (e.g., Divis & Benjamin, 2014). These effects follow from a view that interleaved retrieval serves to spread out the memory representations, making them more distinct from one another. In this dissertation I tested further predictions of retrieval-induced distinction and possible boundaries to the effect.

A key prediction of retrieval-induced distinction is that materials that are mentally spread out via retrieval should have fewer overlapping cues and thus be less likely to interfere with each other. The free recall designs of prior work were not optimal for testing interlist intrusion effects (due to the generally low occurrence of intrusions). Experiments 2A and 2B in the current manuscript overcame the limitations of free recall designs and extended the general finding by showing that interlist confusions are reduced when the lists are separated by a short period of semantic retrieval.

Additional research explored the role of task switching in producing the effects of list distinction. While the typical design controls for the transition out of study by always completing a short control task prior to continuing with that control task or switching to the retrieval task, the task switching seen in the retrieval condition cannot fully account for the effects of retrieval-induced distinction. Spending the entire time doing only the retrieval task or switching between the retrieval and distractor task led to similar effects. It appears that doing *some* of amount of retrieval is what is most critical in this paradigm, supporting the idea that it is the retrieval itself driving the list distinction. That retrieval can be for the entire intervention (e.g., the pure retrieval condition in Experiment 1), for a subset of the intervention (e.g., the switch condition in

Experiment 1; Divis & Benjamin, 2014), for an episodic or semantic retrieval task (e.g., Pastötter et al., 2011), or a difficult or easy retrieval task (e.g., Sahakyan & Hendricks, 2012).

While the critical predictions of retrieval-induced distinction held for basic materials (i.e., lists of words), they started to break down for more complex, narrative materials that were related to one another (i.e., eye witness testimonies of the same crime). Divis and Benjamin (2014) found strong accuracy effects using complex, narrative materials (ACT reading comprehension texts), but in that design each story was about a different topic and did not include overlapping narrative elements. Experiment 3 in the current manuscript (using eye witness testimonies of the same crime) begins to reveal important boundaries on the influence of retrieval-induced distinction.

Based on the eye witness testimony experiment, if the effects of retrieval-induced distinction are to be seen as a viable option for more realistic settings (e.g., the classroom), further work will be needed to determine exactly which aspect led to the predominantly null effects. One prime candidate is that the material was simply too well-learned, leading to ceiling effects and suggesting that retrieval-induced distinction may be most useful for material more weakly learned at the time of the test. Students who are struggling the most are likely to be influenced by these interventions. If it is simply a matter of the difficulty of the material or how well-learned it is, then retrieval-based interventions may still be appropriate in some settings.

Another candidate that warrants further exploration is how connected the narrative elements are across texts. The eye witness testimonies had many elements in common, leading to many shared retrieval cues. Perhaps the interconnectedness inherent in the materials overwhelmed most retrieval-induced distinction effects. The stories were already so connected together that spreading out the memory representations via retrieval could not overcome those overlapping elements. If so, these results point to a critical boundary in the application of retrieval-induced distinction. Only materials that are not inherently strongly tied to each other can benefit from interleaved retrieval. This would suggest that breaking up study of similar topics (e.g., English chapters) with retrieval would not be helpful but that breaking up study of different topics might be (e.g., English and history chapters).

From a more theoretical perspective, future research could also delve into exactly what causes the critical differentiation of mental representations seen in retrieval-induced distinction. Internal context seems to play a key role and accounts for the effects seen in the literature better than explanations based on inhibition or interruption of consolidation interpretations. But where does the line lay between "effective" tasks such as semantic retrieval and "ineffective" control tasks such as counting backward? Prior research indicates it is not simply a matter of task difficulty (Sahakyan & Hendricks, 2012), the match between study and intervening task material (e.g., verbal), or the amount of task switching (Experiment 1). The study material is generally verbal (e.g., lists of words or short texts). While the intervening task material is also often verbal (e.g., semantic retrieval) and the control task often is not (e.g., counting), that difference cannot fully account for the effects. Greater overlap would predict more interference; from that perspective, verbal retrieval events should lead to worse performance overall due to interference. But the effects are not so simple: List 1 performance is reduced and List 5 performance is *enhanced*. Furthermore, retrieval of nonverbal information (such as the n-back task used by Pastötter et al., 2011) led to the same beneficial effect as retrieval of verbal information (semantic or episodic retrieval). Studying verbal information (restudying the previous list in Pastötter et al., 2011) led to the same effect as a control counting condition (reduced performance compared to retrieval). Speeded reading tests (verbal information) do not lead to the same directed forgetting-like effects as events that are posited to induce more context change (Delaney et al., 2010). Experiment 1 in the current manuscript ruled out simple task switching (for semantic retrieval and counting tasks). Work in the directed forgetting literature using context change explanations found that imagination tasks led to effects similar to those of a forget cue but not mathematical tasks (Sahakyan & Kelley, 2002; Pastötter & Bäuml, 2007; Pastötter et al., 2011; Sahakyan, Delaney, & Goodman, 2008), number searches (Mulfi & Bodner, 2010), or speeded reading tasks (Delaney et al., 2010). Where is the line between the retrieval or imagination tasks that appear to lead to more context change and the mathematical, restudying, or reading tasks that do not? Is it the *recall* component of retrieval and imagination that is critical as opposed to the predominantly *recognition* or *procedural* aspect to the other tasks? Further research is needed to draw a more definitive line between effective and ineffective interventions.

Summary

Evidence of retrieval-induced distinction continued to hold for basic materials. Importantly, these studies provided support for the critical prediction of reduced confusions between retrieval-segregated lists in addition to the overall accuracy results found in prior work. While the potential effect of task switching was tested and accounted for, I did find important boundaries to the influence of intervening retrieval that warrant further exploration. When using complex, interconnected material that was well learned, most of the effects of retrieval-induced distinction were extinguished (although there was still evidence for reduced confusions between the materials). Future studies could examine the driving forces behind that boundary or further delve into the theoretical mechanisms (e.g., context change) driving retrieval-induced distinction effects.

References

Bates, D., Maechler, M., Bolkner, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., & Grothendieck, G. (2015). Lme4: Linear mixed-effects models using ‘Eigen’ and S4 [Computer software manual]. Retrieved from http://lme4.r-forge.r-project.org (R package version 1.1-10).

Bjork, E. L. & Bjork, R. A. (1996). Continuing influences of to-be-forgotten information. *Consciousness and cognition*, *5*, 176-196.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavioral Research*, *46*, 904-911.

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

Delaney, P. F., Sahakyan, L., Kelley, C. M., & Zimmerman, C. A. (2010). Remembering to forget: The amnesic effect of daydreaming. *Psychological Science*, *21*(*7*), 1036-1042.

Divis, K. M. & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition*, *42*, 1049-1062.

Earhard, M. (1967). Cued recall and free recall as a function of the number of items per cue. *Journal of Verbal Learning and Verbal Behavior*, *6*, 257-263.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. New York: Dover (original work published 1885).

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*(*3*), 145-154.

Finley, J. R. & Benjamin, A. S. (2012). Adaptive changes in encoding strategy with experience: Evidence from the test expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 632-652.

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.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.), Oxford, U.K.: Oxford University Press

Kahana, M. J. & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin and Review*, *12*(*1*), 159-164.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90(430)*, 773-795.

Mensink, G. & Raaimakers, J. G. (1988). A model of interference and forgetting. *Psychological Review*, *95*(*4*), 434-455.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from http://bayesfactorpcl.r-forge-r-project.org (R package version 0.9.12-2).

Mulji, R. & Bodner, G. E. (2010). Wiping out memories: New support for a mental context change account of directed forgetting. *Memory*, *18*(*7*), 763-773.

Murdock, B. B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology*, *60*(*4*), 222-234.

Nairne, J. S. (2002). The myth of encoding-retrieval match. *Memory*, *10 (5/6)*, 389-395.

Pastötter, B., & Bäuml, K. T. (2007). The crucial role of postcue encoding in directed forgetting and context-dependent forgetting. *Journal of Experimental Psychology: Learning,* *Memory, and Cognition*, *33*(*5*), 977-982.

Pastötter, B., Schicker, S., Niedernhuber, J. & Bäuml, K. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(*2*), 287-297.

Pelli, D. G. (1997). The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442.

R Development Core Team. (2008). R: A language and environment for statistical computer [Computer software manual]. Vienna, Austria. Retrieved from http://www.r-project.org (ISBN3-900051-07-0).

Reitman, W., Malin, J. T., Bjork, R. A., & Higman, B. (1973). Strategy control and directed forgetting. *Journal of Verbal Learning and Verbal Behavior*, *12*, 140-149.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*(*3*), 365-372.

Roediger, H. L., & Guynn, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (Eds.) *Memory*, (pp. 197-236). New York: Academic Press.

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Sahakyan, L., Delaney, P. F., & Goodmon, L. B. (2008). Oh, Honey, I already forgot that:

Strategic control of directed forgetting in older and younger adults. *Psychology and Aging*, *23*(*3*), 621-633.

Sahakyan, L. & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-the-last paradigm. *Memory & Cognition*, *40*(*3*). 844-860.

Sahakyan, L. & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology, 28*(*6*), 1064-1072.

Sahakyan, L. & Smith, J. R. (2013). "A long, long ago, in a context far, far away": Retrospective time estimates and internal context change. *Journal of Experimental Psychology: Learning, Memory & Cognition*.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic simulation of future events: Concepts, data, and applications. *Annals of the New York Academy of Sciences*, *1124*, 39-60.

Shiffrin, R. M. (1970). Forgetting: Trace erosions or retrieval failure? *Science*, *168*(*3939*), 1601-1603.

Snodgrass, J. G. & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(*1*), 34-50.

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

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. *Journal of Experimental Psychology: General*, *143(4)*, 1526-1540.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80(5)*, 352-373.

Tulving, E. & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior, 13*, 181-193.

Wetzels, R., Matske, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6(3)*, 291-298.

Appendix A: Mixed Effects Model Fitting

Significance levels for all statistical tests were set at an $\alpha < .05$ level. Random effects models were fit using maximum likelihood estimation and Satterthwaite approximations to degrees of freedom in the lme4 package (Bates et al., 2011) in R software (R Development Core Team, 2009). Binomial data (e.g., accuracy) was fit using Laplace estimation.

The best model was chosen by including the fixed effects of interest and then determining whether including random effects significantly improved the fit of the model for the increase in complexity. Models that were *not nested* were compared using the Akaike information criterion (*AIC*) and Bayesian information criteria (*BIC*) values. *Nested* models were compared using a Chi-square test with the *anova()* function in R software to evaluate whether the reduction in residual sum of squares between the two models was statistically significant (consistent with reporting standards, this statistic is only reported for significant improvements in fit of the model; $p < .05$). The model with the best fit was then chosen for further analysis. Significant test statistic values ($p < .05$) are always reported when comparing *nested models*; *AIC* values (for *non-nested models*) are reported for the top two non-nested candidates at the end of each model fitting section.

Experiment 1

Proportion correct. The fixed effects of interest were intervening task (pure retrieval, pure distractor, or switching) and list tested (List 1 or List 5); all models included these fixed effects. The random effects of interest were subject and item. Model 1 included random intercepts for subject; Model 2 included random intercepts for item. Model 2 led to better *AIC* and *BIC* values. Model 3 included random intercepts for both subject and item; it did not fit the data significantly better than Model 2. Model 4 included random slopes for item; it failed to converge, indicating it was not an appropriate model for this data. Overall, Model 2 fit the data best (*AIC* = 1587 relative to *AIC* = 1646.3 for Model 1).

Intrusions. The fixed effects of interest were intervening task (pure retrieval, pure distractor, or switching) and list tested (List 1 or List 5); the random effect of interest was subject. Model 1 included the fixed effects and random intercepts for subject. Given the structure of the data, it is not appropriate to include random slopes for subject (the model would be unidentifiable). Model 1 fit the data best.

Experiment 2A

Near-list intrusions on List 3. The fixed effects of interest were intervening task (retrieval or distractor) and intrusion source list (List 2 or List 4); all models included these fixed effects. The main potential random effect of interest was subject; the secondary potential random effects of interest were proportion wrong on intrusion source list and proportion wrong on List 3. Model 1 included random intercepts for subject. Given the structure of the data, it is not appropriate to include random slopes for subject (the model would be unidentifiable). Model 2 included random intercepts for both subject and performance on intrusion source list; it fit the data significantly better than Model 1 ($\chi^2(1)$ = 11.857, p < .001). Model 3 included random intercepts for both subject and performance on List 3; it did not fit the data better than Model 1 and had worse *AIC* and *BIC* values than Model 2. Model 4 included random intercepts for subject, performance on the intrusion source list, and performance on List 3; it did not fit the data significantly better than Model 2. Model 5 updated Model 2 to include random slopes for performance on the intrusion source list; it failed to converge, indicating it was not an appropriate model for this data. Overall, Model 2 fit the data best (*AIC* = -50.12 relative to *AIC* = -41.096 for Model 3).

All near-list intrusions. The fixed effect of interest was intervening task (retrieval or distractor); the random effect of interest was subject. Model 1 included the fixed effect of task and random intercepts for subject. Model 2 included the fixed effect of task and random slopes for subject. Model 2 fit the data significantly better than Model 1 ($\chi^2(2)$ = 6.560, $p < .037$), indicating it was the best model for this data.

Accuracy on List 1. The fixed effect of interest was intervening task (retrieval or distractor); all models included this fixed effect. The potential random effects of interest were subject and item. Model 1 included random intercepts for subject; Model 2 included random intercepts for item. Model 1 led to better *AIC* and *BIC* values. Model 3 included random intercepts for both subject and item; it did not fit the data significantly better than Model 1. Model 4 included random slopes for subject; it also did not fit the data significantly better than Model 1. Overall, Model 1 fit the data best (*AIC* = 377.33 relative to *AIC* = 419.24 for Model 2).

Experiment 2B

Interlist intrusions on List 2. The fixed effects of interest were intervening task (retrieval or distractor) and intrusion source list (List 1 or List 3). The random effect of interest was subject. Model 1 included the fixed effects and random intercepts for subject. Given the structure of the data, it is not appropriate to include random slopes for subject (the model would be unidentifiable). Model 1 fit the data best.

Accuracy on List 2. The fixed effect of interest was intervening task (retrieval or distractor); it was included in all models. The potential random effects of interest were subject and item. Model 1 included random intercepts for subject; Model 2 included random intercepts for item. Model 1 had better *AIC* and *BIC* values relative to Model 2. Model 3 included random intercepts for both subject and item; it did not fit the data significantly better than Model 1. Model 4 included random slopes for subject; it failed to converge, indicating it was not an appropriate model for the data. Overall, Model 1 fit the data best (*AIC* = 840.07 relative to *AIC* = 854.91 for Model 2).

Experiment 3

Accuracy: Multiple Choice. The fixed effects of interest were intervening task (retrieval or distractor) and test position (first or last); all models had these fixed effects included. The main potential random effects of interest were subject and question; secondary random effects of interest were eye witness tested (which is also tied to question) and eye witness testimony order (which is also tied to test position). Model 1 included random intercepts for subject; Model 2 included random intercepts for question. Model 2 had better *AIC* and *BIC* values. Model 3 included random intercepts for both subject and question. When compared to Model 2, Model 3 was a significantly better fit ($\chi^2(1)$ = 19.796, $p < .001$). Model 4 included random slopes for question and random intercepts for subject. Model 5 included random slopes for subject and random intercepts for question. Both Model 4 and Model 5 failed to converge, indicating they did not fit the data well. In addition to the effects in Model 3, Model 6 included random intercepts for eye witness tested. In addition to the effects in Model 3, Model 7 included random intercepts for eye witness testimony order. Neither Model 6 nor Model 7 led to significantly better model fits compared to Model 3. Overall, Model 3 fit the data best[23].

[23] All other models were nested versions of this model and could be directly tested, so I do not report the *AIC* values here.

Foil Endorsement: Multiple Choice (Contradictory). The fixed effects of interest were intervening task (retrieval or distractor) and question; all models included these fixed effects. Once again, the main potential random effects of interest were subject and question; secondary random effects of interest were eye witness tested (which is also tied to question) and eye witness testimony order (which is also tied to test position). Model 1 included random intercepts for subject; Model 2 included random intercepts for question. Model 2 led to better *AIC* and *BIC* values. Model 3 included random intercepts for both subject and question. When compared to Model 2, Model 3 did not lead to a significantly better fit. Model 4 included random slopes for question; it failed to converge, indicating it was not an appropriate model for the data. Model 5 included random intercepts for question and eye witness tested; Model 6 included random intercepts for question and eye witness order. Neither Model 5 nor Model 6 fit the data significantly better than Model 2. Overall, Model 2 fit the data best (*AIC* = 588.31 relative to *AIC* = 754.13 for Model 1).

Misattributions: Individuals (All Data). The fixed effect of interest was intervening task (retrieval or distractor); it was included in all models. The main potential random effects of interest were subject and question; secondary random effects of interest were eye witness tested in the multiple choice test, position tested in the multiple choice test, and eye witness order. Model 1 included random intercepts for subject; Model 2 included random intercepts for question. Model 2 led to better *AIC* and *BIC* values. Model 3 included random intercepts for both subject and question; it overfit the data for repeated information, indicating it was not an appropriate model for the data. Model 4 included random slopes for subject and random intercepts for question; it failed to converge, indicating it was not a good match for the data. Model 5 included random intercepts for subject and random slopes for question; when compared to Model 2, it did not fit the data significantly better. Model 6 included random intercepts for subject, question, and eye witness tested in the multiple choice test. Model 7 included random intercepts for subject, question, and position tested in the multiple choice test. Model 8 included random intercepts for subject, question, and eye witness order. Models 6-8 did not fit the data significantly better than Model 2. Overall, Model 2 fit the data best (*AIC* = 8700.8 relative to *AIC* = 8895.6 for Model 1).

Misattributions: Individuals (Untested Data). Once again, the fixed effect of interest was intervening task (retrieval or distractor), and it was included in every model. The main potential random effects of interest were subject and question; secondary random effects of interest were eye witness tested in the multiple choice test, eye witness position tested in the multiple choice test, and eye witness order. Model 1 included random intercepts for subject. Model 2 included random intercepts for question. Model 2 had better AIC and BIC values compared to Model 1. Model 3 included random intercepts for both subject and question; it fit the data significantly better than Model 2 ($\chi^2(1)$ = 167.43, p < .001). Model 4 included random slopes for subject and random intercepts for question; it failed to converge, indicating it was not an appropriate model for the data. Model 5 included random intercepts for subject and random slopes for question; it did not fit the data significantly better than Model 3. Model 6 included random intercepts for subject, question, and eye witness tested in the multiple choice test. Model 7 included random intercepts for subject, question, and position tested in the multiple choice test. Model 8 included random intercepts for subject, question, and eye witness order. Models 6-8 did not fit the data significantly better than Model 3. Overall, Model 3 fit the data best[24].

[24] Once again, this model was a nested version of all other models and could be directly tested so I do not report the *AIC* values here.

Appendix B: Bayes Factors

The Bayes factor provides a simple, intuitive method for determining the degree to which data support the null or alternative hypothesis (see Jeffreys, 1961; Kass & Rafety, 1995). A larger Bayes factor (B_{10}) indicates stronger evidence for the alternative hypothesis. Given the prior odds are assumed to be 1.0, a Bayes factor (B_{10}) of 3.0 means the data were three times as likely to have occurred under the alternative hypothesis as the null hypothesis. See Table 2 for the guidelines Jeffreys (1961) and Kass & Rafety (1995) provide for interpreting Bayes factors. Notably, these guidelines were intended for large sample sizes. Morey et al. (2009) simulated small effect sizes across a range of sample sizes and showed that for small to moderate samples (<5,000 in their simulation), the Bayes factor supports the null hypothesis (at larger sample sizes, it supports the alternative hypothesis).

Mixed effects models were intentionally chosen for the main analyses because they best match the needs of this project. Both subject and item effects can be accounted for, soaking up some of the unexplained variance that is left in more traditional measures (e.g., classic t-test). Using mixed effects models is also the most consistent with prior work (Divis & Benjamin, 2014). One limitation of using a complex method is that it does not always fit with other methods. To the best of my knowledge, tools for calculating Bayes factors based on test statistics generated by mixed effects models have yet to be created.[25] That means in order to calculate Bayes factors for the data in the current experiments, I must regress back to simpler t-tests (and lose the significant variance accounted for by the mixed effects models). These Bayes factors,

[25] Wetzels and colleagues (2011) compared p-values, effect sizes, and Bayes factors for 855 t-tests in published psychology studies. P-values from .01 to .05 tended to be associated with Bayes factors that indicated weak or anecdotal evidence for the alternative hypothesis. P-values less than .01 were associated with Bayes factors that indicated substantial evidence for the alternative hypothesis. If one were to extrapolate that this same pattern would hold for mixed effects model test statistics, p-value thresholds of .05 and .01 could be used in a similar way to determine the level of evidence in support of the alternative hypothesis.

based on simpler t-tests, are still valuable but should be interpreted with full knowledge that they are not directly drawn from the mixed effects model results reported in the main text. Notably, with the small sample sizes used in these experiments, Bayes factors are more likely to indicate evidence for the null hypothesis for small effect sizes.

Bayes factors were calculated using the BayesFactor package (Morey, Rouder, & Jamil, 2015) in R software, using Student's t-test statistics (*not* the mixed effects model test statistics). The Bayes factors were computed via Gaussian quadrature, using a medium (r = .707) sized prior (see Rouder, Speckman, Sun, Morey, & Iverson, 2009). Tables 3-6 report the mixed effects model test statistics, Student's t-test statistics, Bayes factors, and interpretations for Experiments 1, 2A, 2B, and 3, respectively.

Appendix C: Experiment 3 Materials

Eye Witness Testimonies

<u>Tony the Retired Paramedic</u>

That afternoon, I was sitting on a bench at the park while my wife was shopping for a new dress in the stores across the street. When the weather is nice, I prefer to sit at the park people watching, rather than go shopping with her. My bench was at the corner of the park, near a hot dog stand. It was interesting to see all the people coming and going in the park. There were businessmen, joggers, families, retirees, dog walkers—all types of people. I noticed Becky and her husband right away—with their maps and excited but confused expressions they stood out as tourists exploring the city.

I remember thinking it was a bad idea for Becky to flash all that cash when she paid for the hot dogs. Sure enough, someone decided to take advantage of that opportunity. A man that had been hanging out around the stand came up behind her, shoved her to the ground and took off out of the park with her purse. I saw the flash of a small knife—he must have cut the strap of the purse to steal it. I'm a retired paramedic, so I immediately ran to Becky to make sure she was okay rather than trying to chase off after the thief. I identified myself and asked what her name was. She was pretty dazed and confused. She had a 3 centimeter abrasion on her forehead, near her hairline. It was swollen with minor bleeding, but thankfully there weren't any signs of a concussion. She also had torn her pants and had small abrasions on her hands and knees from her fall.

I had noticed the assailant standing around the hot dog stand before the crime. The vendor had asked him if he wanted a hot dog and he had rudely declined. His voice sounded like he might be a smoker—kind of gravelly. He was average height, somewhere around 5'9" and had on a t-shirt and a cap with a Chicago Cubs logo. He also had a dark brown jacket tied around his waist. He had a darker skin tone and looked like he might be of Pakistani descent. It looked like he hadn't shaved for a few days. I remember a large mole on the left side of his face that I thought should be checked out by a dermatologist. There were also tattoos going up and down both arms.

Gary the Hot Dog Vendor

That day I had my hot dog stand set up in my usual spot—at the corner of Willis Park near an ice cream vendor, right across the street from Macy's and Wells Fargo Bank. In the middle of the lunchtime rush I noticed a man hanging around my stand watching me and my customers. I thought maybe he was just deciding whether or not he wanted to buy a hot dog, so I called out to him with my usual spiel. He made an obscene gesture, telling me that he wasn't going to buy a hot dog and that I should mind my own business. I see those rude types all the time, so I decided to just ignore him and focus on the customers waiting in line.

About 10 minutes later, a middle aged woman and her husband ordered two of my lunch specials. After I told her the total cost, she reached into her purse and pulled out a wallet with a bunch of cash. She had just put her wallet back into her purse when all the sudden the man I had noticed hanging around my stand earlier came up behind her, attacked her, and stole her purse. He pushed her to the ground and pulled the purse so hard that the strap broke. He ran off behind me with the purse. It looked like he was heading toward the business district. The woman's husband took off after the thief. The woman looked dazed and had a cut and large bump on her forehead. A man who identified himself as a retried paramedic came over to look after her.

I was able to get a good look at the suspect before the robbery occurred. He was about my height—I'd say between 5'8" and 5'10". He had darker skin and looked like he might be of Indian descent. He had on a white baseball cap, so I couldn't see his hair color. He was wearing jeans and a short-sleeved polo shirt with a black jacket tied around his waist. I noticed a large mole on his cheek below his eye and tattoos on both of his arms. He had some stubble on his face but not a full beard. Before the crime, when I asked him if he wanted to order a hot dog, I noticed that he had a deep, low-pitched voice.

Chris the Victim's Husband

My wife, Becky, and I were touring the city as part of a road trip from Milwaukee to Nashville we were taking this year. We had just finished an architectural walking tour nearby and decided to stop at a hot dog stand for lunch. After ordering Becky pulled out some cash and paid for the meal. She had just put her wallet back in her purse and was trying to zip the purse closed when suddenly out of nowhere a man came up behind us, shoved her to the ground, and took off with her purse. Her purse had all of the important things for the trip—the keys to the hotel and rental car, money for the trip, and our reservations. My instincts kicked in, and I immediately chased after the thief.

He ran out of the park and across the street toward a shopping area and business district. I was just about to cross the street to pursue him when I saw him get into the passenger side of a car parked at the end of the block. The car took off down the street, zooming right past me before turning north a few blocks later. It was a black, midsized car that looked like an older model. I think it might have been a Toyota Camry. It was junky looking—there were patches of missing paint on the trunk and the muffler was hanging down. It also had a Chicago Bears license plate.

All I saw of the man who pushed my wife and stole her purse was his back as I was chasing him. He was wearing a short-sleeved shirt, jeans, and a baseball cap. However, I was able to get a look at the driver as the car drove by me. The driver looked like a teenage boy. He was white and had dark brown hair with a shaggy haircut. He was wearing dark sunglasses and a red hoodie.

Steven the Businessman

I work at Charles Schwab Investments near Willis Park. The day of the crime, I had spent the entire morning trying to fix a critical mistake one of our new employees made. I wanted to get out of the office, so I decided to take my lunch to the park. I was just finishing up my ham sandwich and considering treating myself to an ice cream cone when I noticed a commotion over by the hot dog stand. A middle-aged woman was on the ground, and a man wearing a light colored cap and short-sleeved shirt was running toward me with a purse in his hands. It looked like he had just stolen the lady's purse.

The man sprinted right past me and across the street toward the commercial district. On the way out of the park, he ran right through a muddy area. His shoe print might still be there. Once the thief got across the street, he went about 50 feet down the sidewalk and got into a car near the Starbucks. There must have been a driver waiting in the car--the thief didn't even have the passenger door all the way closed before the car was peeling out of its parking spot and heading back my way. I got a good look at the car and was also able to make out the driver before it turned onto Washington Avenue.

The driver looked like a young, Caucasian woman. She was wearing a red sweatshirt with the hood partially pulled up and dark sunglasses. She had black, medium length hair. The car was dark blue but wasn't in very good condition—the windshield was cracked and the bumper was rusted out in places. It looked like a Honda Accord, and I'd say it was about 10 years old. I wasn't able to catch the entire license plate but it started with "G97."

Multiple Choice Questions

Correct answers are in bold; foils are in red.

Tony and Gary (Tony Tested)

Contradictory Information:

1.What happened to the victim's purse when the thief took it?

a. The thief cut the strap with a knife.
b. The thief pulled it so hard that he broke the strap.
c. The thief tore a hole in the side of the purse when he ripped it out of the victim's hands.
d. The strap got tangled in the victim's hair.

2.What ethnicity was the thief?

a. Indian
b. Pakistani
c. Turkish
d. Cambodian

3.What type of shirt was the thief wearing?

a. Short-sleeved polo shirt
b. T-shirt
c. Jersey
d. Workout shirt

4.What color was the thief's jacket?

a. Black
b. Brown
c. Blue
d. Gray

Superadditive Information:

5.How did Tony describe the thief's hat?

a. White baseball cap

b. Had a Chicago Cubs logo

c. Still had a sticker on the bill

d. Dirty and grimy

6.Where was the large mole on the thief's face?

a. On his cheek below his eye

b. On the left side of his face

c. Near the corner of his eye

d. On his nose

7.How did Tony describe the thief's voice?

a. Deep, low-pitched

b. Gravelly

c. Soft

d. Fast-paced

Repeated Information:

8.What was the victim's mental state immediately after the crime?

a. Dazed

b. Angry

c. Concerned

d. Embarrassed

9.What injuries did the victim sustain on her head?

a. Bleeding laceration on cheek

b. Minor bleeding from abrasion on forehead

c. Large contusion on back of head

d. Eye swollen shut

10. How tall was the thief?

a. Around 5'9"

b. Around 5'11"

c. Around 6'1"

d. Around 6'3"

11. How was the thief wearing his jacket?
    - **a. Tied around his waist**
    - b. With the sleeves rolled up
    - c. Zipped halfway up
    - d. Instead out

12. Where were the thief's tattoos located?
    - **a. On both arms**
    - b. Only on the left arm
    - c. Only on the right arm
    - d. He didn't have tattoos

13. What sort of facial hair did the thief have?
    - a. None--he was clean shaven
    - b. Full beard
    - **c. Hadn't shaved for a couple days**
    - d. Messy goatee

Unique Information:

14. What was Tony's wife shopping for while he was at the park?
    - **a. Dress**
    - b. Jewelry
    - c. Shoes
    - d. Jeans

15. What did Tony say he liked to do at the park while his wife shopped?
    - **a. People watch**
    - b. Take a nap
    - c. Eat lunch
    - d. Go for a walk

16. What type of people did Tony **NOT** mention seeing at the park?
    - a. Joggers
    - b. Dog walkers
    - c. Businessmen
    - **d. College students**

17. Which of the following did Tony say the victim had?
    a. Concussion
    b. Small abrasions on her hands and knees
    c. Sprained ankle
    d. Heart palpitations

Steven and Chris (Steven Tested)

Contradictory Information:

1. What color was the getaway car?
    a. Black
    b. Dark blue
    c. Brown
    d. Dark green

2. What make and model was the getaway car?
    a. Honda Accord
    b. Toyota Camry
    c. Hyundai Sonata
    d. Nissan Altima

3. What was the gender and age of the driver?
    a. Teenage boy
    b. Young woman
    c. Teenage girl
    d. Young man

4. What color was the driver's hair?
    a. Dark brown
    b. Black
    c. Auburn
    d. Dark blonde

Superadditive Information:

5.How did Steven describe the route of the getaway car?
    a. Turned north a few blocks after driving past him.
    b. Turned onto Washington Avenue
    c. Turned at the second traffic light down the street
    d. Made a U-turn and went east

6.What was the state of the getaway car?

a. Junky with patches of missing paint on the trunk
b. Poor condition with a cracked windshield
c. A junker with a rusted out bumper
d. Pretty beaten up with a large dent in the door

7.What did the license plate of the getaway car look like?

a. Chicago Bears specialty plate
b. Started with "G97"
c. Ended with "352"
d. University of Illinois specialty plate

Repeated Information:

8.Where was the getaway car originally parked?

a. Across the street near a Starbucks
b. On the other side of the park near an ice cream stand
c. Out of sight on the other side of the business district
d. Near a garbage truck across the street

9.How old was the getaway car?

a. Brand new
b. A couple years old
c. About a decade old
d. Over 25 years old

10. What was the driver's ethnicity?

a. Caucasian
b. Asian
c. African American
d. Hispanic

11. How did Steven describe the driver's hair?

a. Short cut
b. Medium length cut
c. Curly
d. Frizzy

12. What accessory was the driver wearing?
    a. **Dark sunglasses**
    b. Large ring
    c. Diamond stud
    d. Scarf

13. What color was the driver's shirt?
    a. **Red**
    b. Orange
    c. Yellow
    d. Green

Unique Information:

14. Where does Steven work?
    a. **Charles Schwab Investments**
    b. TD Ameritrade
    c. Merrill Edge
    d. Fidelity Investments

15. What did Steven spend the morning doing?
    a. Interviewing job candidates
    b. Meeting with a new client
    c. **Fixing another employee's mistake**
    d. Organizing a presentation

16. What did Steven eat for work?
    a. **Ham sandwich**
    b. BLT
    c. Burrito
    d. Pizza

17. What did Steven say about the thief's shoes?
    a. **He might have left shoe prints in a muddy patch.**
    b. He lost a shoe while running across the park.
    c. They were soaked from running through a puddle.
    d. They were worn out and almost falling off his feet while he ran.

Example Short Answer Test Booklet (Matrix Portion)

For this part of the experiment, your task is to answer the following questions about the crime and which eye witness(es) provided that information. Please complete this worksheet using the following steps:

1. Answer the question, using information from ALL FOUR eye witness accounts. It's okay if there is contradictory information—please write down everything you remember reading about that question.
2. Indicate which eye witness provided which information. If an eye witness provided NO information on that topic, put an "X" in his box.

Example:

Let's pretend Tony said the victim was wearing a red blouse but Gary said she had a pink shirt with a small floral pattern (Chris and Steven did not mention the victim's shirt). Let's also pretend that Chris said she ordered a Pepsi and Gary said she ordered a medium Pepsi (but Tony and Steven didn't mention her drink). You would fill in the table like this:

Category	**Question**	**Answer based on ALL eye witness accounts.**	**Who provided that information?**			
			Tony the Retired Paramedic	Gary the Hot Dog Vendor	Chris the Victim's Husband	Steven the Businessman
Information about the victim.	What did the victim's shirt look like?	*Red or pink blouse with a small floral pattern.*	*Red blouse*	*Pink with a small floral pattern*	*X*	*X*
	What type of soda did the victim order prior to the crime?	*Medium Pepsi*	*X*	*Medium Pepsi*	*Pepsi*	*X*

If you have any questions, please ask the experimenter now. Otherwise, please fill in the table starting on the next page.

Category	**Question**	**Answer based on ALL eye witness accounts.**	**Who provided that information?**			
			Tony the Retired Paramedic	Gary the Hot Dog Vendor	Chris the Victim's Husband	Steven the Businessman
Information about the thief	How tall was the thief?					
	What was the ethnicity of the thief?					
	Describe the thief's voice					
	Describe the thief's facial hair.					
	What did the thief's hat look like?					
	Describe the thief's shirt (including color and type of shirt).					
	Describe the thief's jacket (including color and location).					
	Where were the thief's tattoos located?					

		(Answer based on ALL eye witnesses)	**(Tony)**	**(Gary)**	**(Chris)**	**(Steven)**
	Where on his face was the thief’s mole located?					
Information about what happened to the victim	How did the purse strap break when the thief stole it?					
	What was the victim’s mental state after the crime?					
	Describe the victim’s physical injuries after the crime.					
Information about the getaway car	What color was the car?					
	What was the make and model of the car?					
	How old was the car?					
	Describe the condition of the car & any problems with it (e.g., missing hubcap)					
	Describe the car’s license plate.					

		(Answer based on ALL eye witnesses)	**(Tony)**	**(Gary)**	**(Chris)**	**(Steven)**
	Where was the car parked before the thief got into it?					
	Describe the getaway route. Where did the car go after passing by the eye witnesses?					
Information about the driver of the getaway car	What was the driver's age and gender					
	What was the driver's ethnicity?					
	Describe the driver's hair (including color and cut).					
	What color was the driver's hoodie?					
	What accessories did the driver have on?					

Short answer test booklet questions on unique information

Tony the Retired Paramedic

1. What did Tony like to do at the park?
2. What was Tony's wife shopping for while he was at the park?
3. What types of people did Tony describe seeing at the park?

Gary the Hot Dog Vendor

1. Where was the hot dog stand located? (Be sure to include which part of the park and the name of businesses across the street).
2. What did the thief do when Gary asked him if he wanted a hot dog?
3. How long was it between Gary asking the thief if he wanted a hot dog and the thief stealing the victim's purse?

Chris the Victim's Husband

1. What cities were the start and end points of Chris' and Becky's road trip?
2. What type of tour were Chris and Becky on before going to the hot dog stand?
3. What important things were in the stolen purse? (Be specific).

Steven the Businessman

1. What's the name of the business where Steven works?
2. What did Steven spend the morning doing?
3. What did Steven eat for lunch?

Appendix D: Experiment 3 Supplementary Analyses

Superadditive Information

All of the analyses reported in the main body of the manuscript (refer to Experiment 3 results) specific to type of information were also conducted on superadditive information using the same mixed-effects models. See Figure 15. No effect of intervening task (retrieval versus distractor) was found for accuracy in the multiple choice test for the simple effect of task for either the first or last testimony. Similarly, intervening task did not influence the overall effect, simple effect for the first testimony, or simple effect for the last testimony for foil endorsement in the multiple choice test. There was also no difference in rate of misattributions between conditions in the short answer test when using either the full dataset or just information not previously tested in the multiple choice test.

Short Answer Test: Accuracy

One of the components of the matrix portion of the short answer test was answering each question based on all of the eye witness testimonies. The mixed effects model that best fit proportion correct on these responses included fixed effects for intervening task (retrieval or distractor) and eye witness pair (either the first or last pair of eye witness testimonies studied), along with random intercepts for subject and random slopes for question. When including all types of information, there was no simple effect of task for either the first or last pair of eye witnesses. This pattern held for both repeated and superadditive information, along with the simple effect of task for the first pair of eye witnesses when looking at only contradictory information. However, the model revealed a simple effect of task for the last pair of eye witnesses for contradictory information, where participants in the retrieval condition had overall lower performance than those in the distractor condition (*t*(27.59) = 2.143, *p* = .041). See Figure 16.

The other component of the matrix portion of the short answer test was identifying which eye witness said what for each question (individual source memory). Accuracy was analyzed by looking at output bound accuracy (when participants indicated the eye witness said something about that question, how often were they correct?). The mixed effects model that best fit proportion correct included fixed effects for intervening task (retrieval or distractor) and eye witness pair, along with random intercepts for subject and random slopes for question. The model revealed no significant effect of task on accuracy when looking at either the simple effect of the first pair of eye witness testimonies studied or the simple effect of the last pair. This pattern held when collapsing across all types of information or breaking it down by repeated information only, contradictory information only, or superadditive information only. See Figure 17.

The last section of the short answer test queried the unique information said by each eye witness. The mixed effects model predicting accuracy that best fit the data included fixed effects of intervening task and eye witness study position, along with random intercepts for subject and question. It revealed no effect of intervening task for either the simple effect of the first eye witness studied or the last eye witness studied. Notably this test was the farthest away from the study session and participants' memory had been queried two times prior to starting this section (in the multiple choice test and the matrix portion of the short answer test). See Figure 18.

Figures

Figure 1

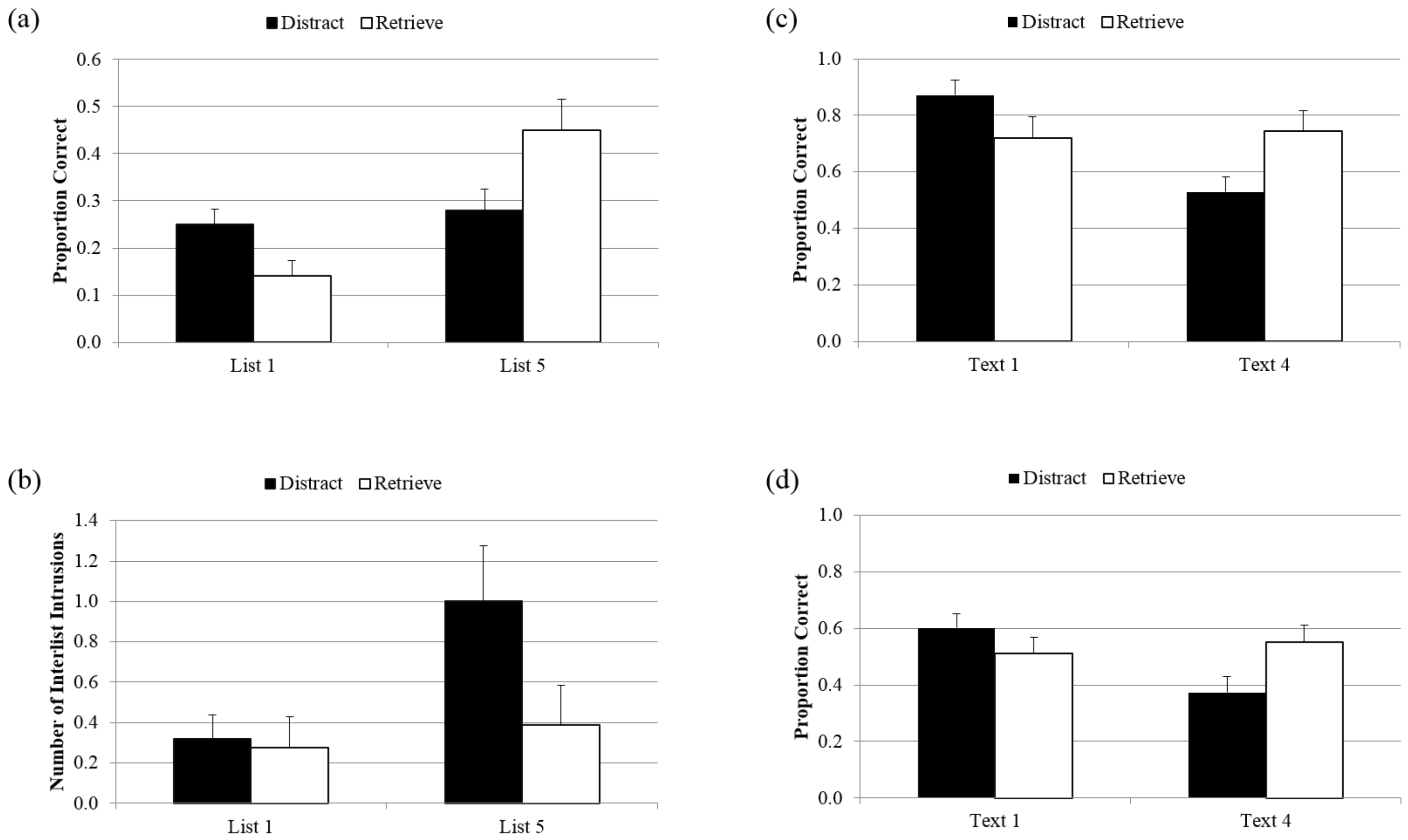


Figure 1. Results from Divis and Benjamin (2014) for (a) proportion correct as a function of list tested (List 1 or List 5) and intervening task (distract or retrieve) for Experiment 1; (b) interlist intrusions as a function of list tested (List 1 or List 5) and intervening task (distract or retrieve) for Experiment 1; and proportion correct for (c) multiple choice and (d) short answer questions as a function of text tested (Text 1 or Text 4) and intervening task (distract or retrieve) for Experiment 2. Error bars represent standard error of the mean.

Figure 2

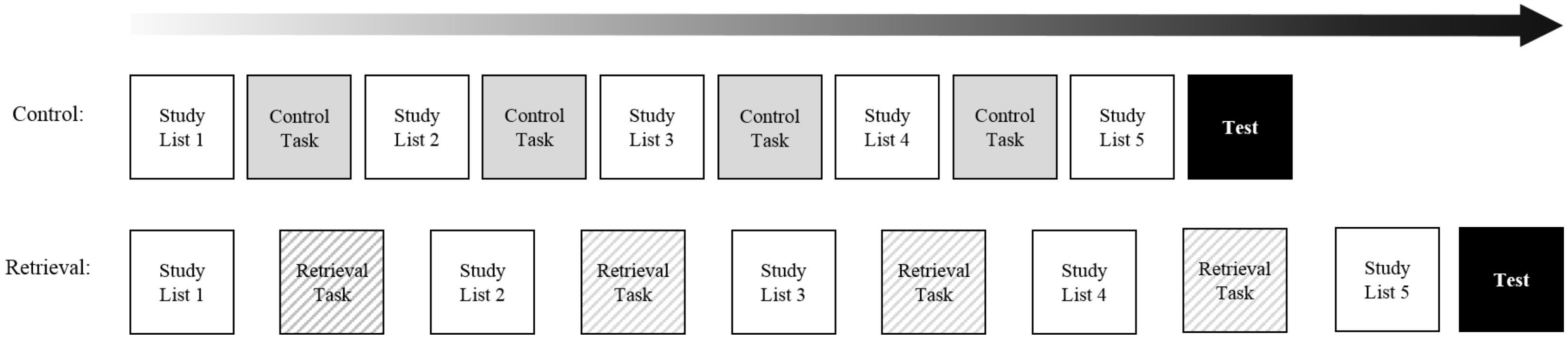
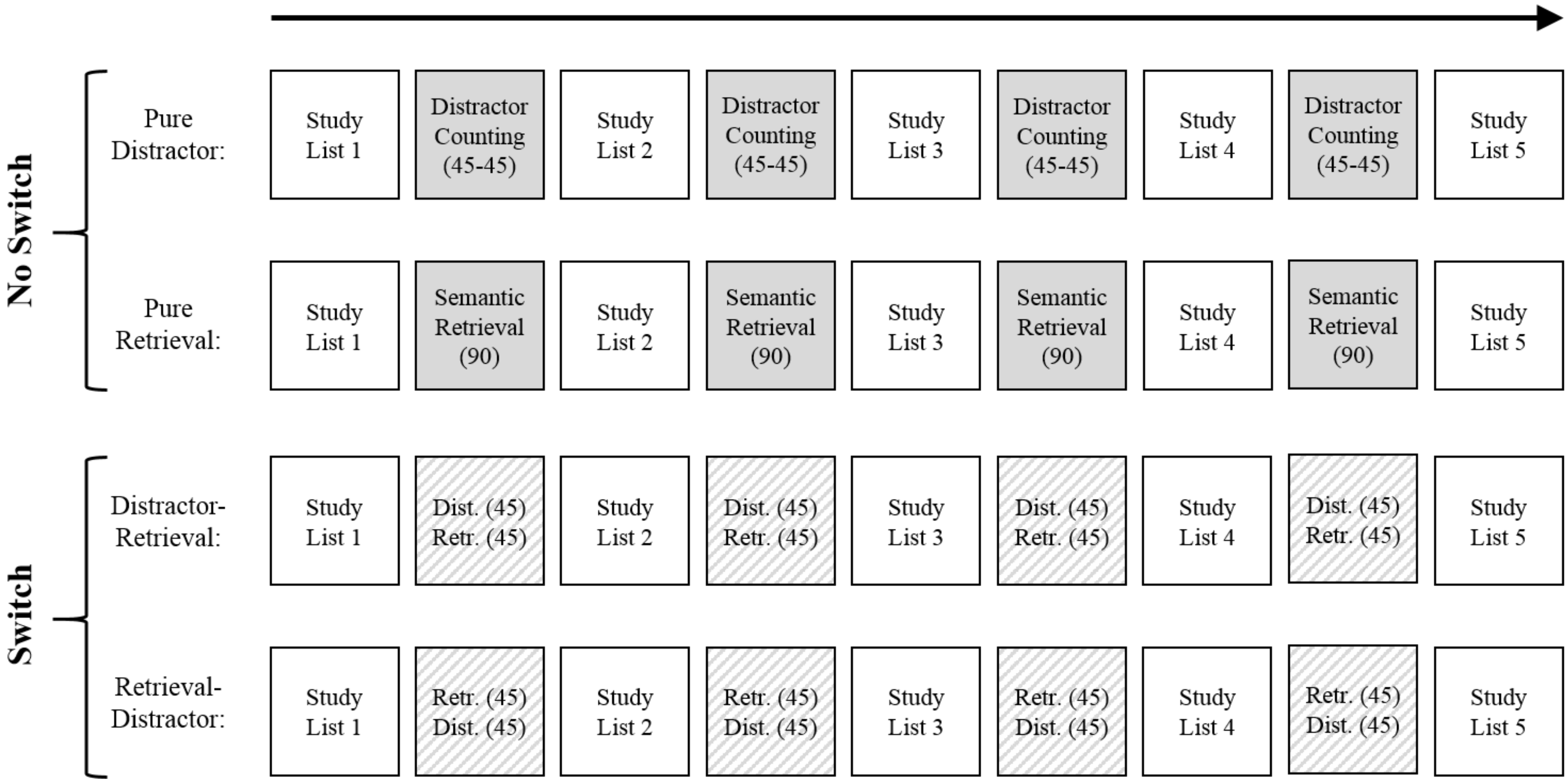


Figure 2. Depiction of the predicted influence of spreading out of mental representations due to interleaved retrieval within a multilist learning paradigm (i.e., retrieval-induced distinction). The gradient arrow represents mental distance (i.e., placement in contextual stream); items farther apart are more dissimilar.

Figure 3



No Switch	Pure Distractor:	Study List 1	Distractor Counting (45-45)	Study List 2	Distractor Counting (45-45)	Study List 3	Distractor Counting (45-45)	Study List 4	Distractor Counting (45-45)	Study List 5
	Pure Retrieval:	Study List 1	Semantic Retrieval (90)	Study List 2	Semantic Retrieval (90)	Study List 3	Semantic Retrieval (90)	Study List 4	Semantic Retrieval (90)	Study List 5
Switch	Distractor-Retrieval:	Study List 1	Dist. (45) Retr. (45)	Study List 2	Dist. (45) Retr. (45)	Study List 3	Dist. (45) Retr. (45)	Study List 4	Dist. (45) Retr. (45)	Study List 5
	Retrieval-Distractor:	Study List 1	Retr. (45) Dist. (45)	Study List 2	Retr. (45) Dist. (45)	Study List 3	Retr. (45) Dist. (45)	Study List 4	Retr. (45) Dist. (45)	Study List 5

Figure 3. Study phase design in Experiment 1 by condition (no switch: 90 seconds of retrieval or distractor; switch: 45 seconds each of distractor and retrieval, order counterbalanced).

Figure 4

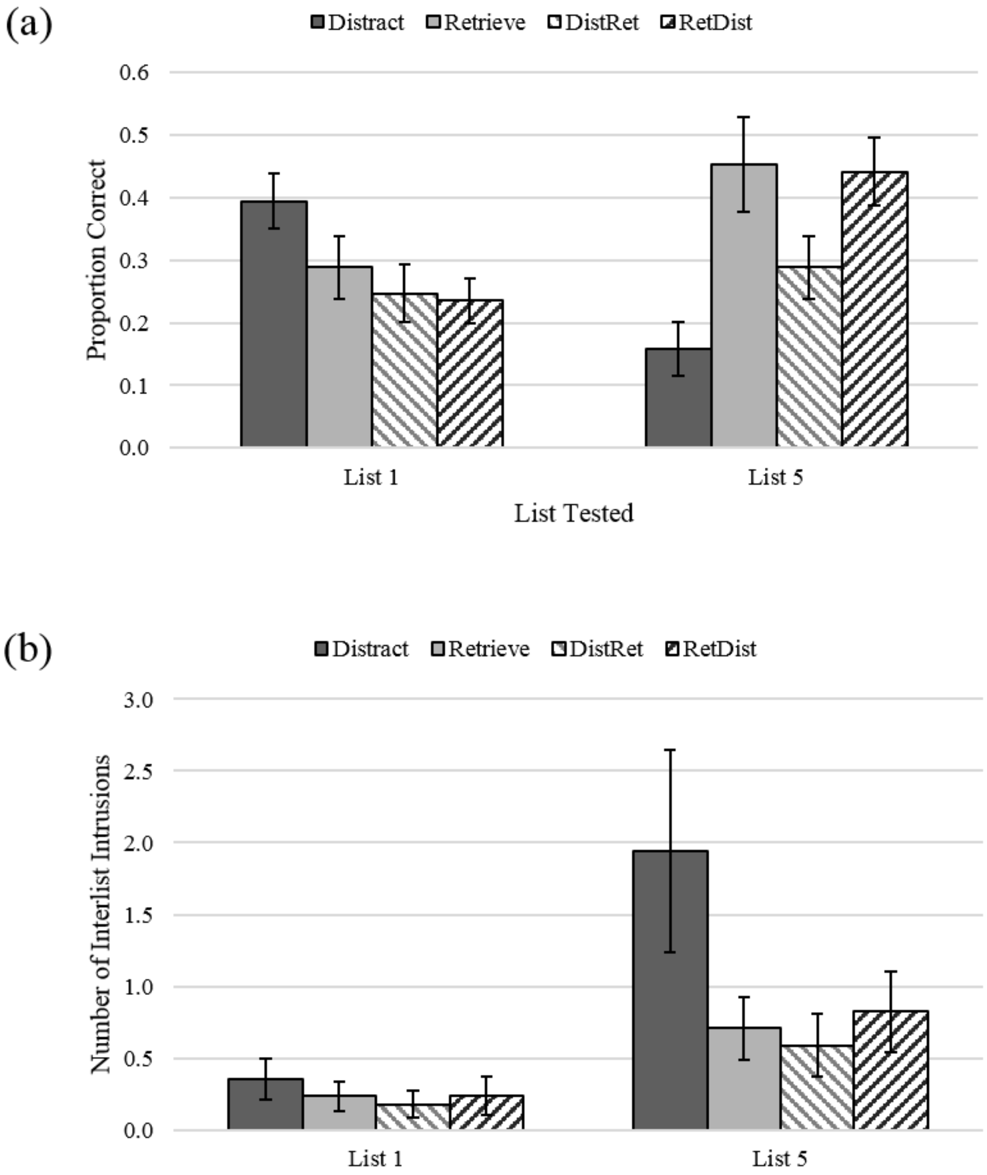


Figure 4. (a) Proportion correct and *(b)* number of interlist intrusions as a function of list (1 or 5) and task (pure distractor, pure retrieval, distract-retrieve, or retrieve-distract) for Experiment 1. Solid bars represent no switch conditions (pure distractor and pure retrieval); striped bars represent switch conditions (distract-retrieve and retrieve-distract). The error bars represent the standard error of the mean.

Figure 5

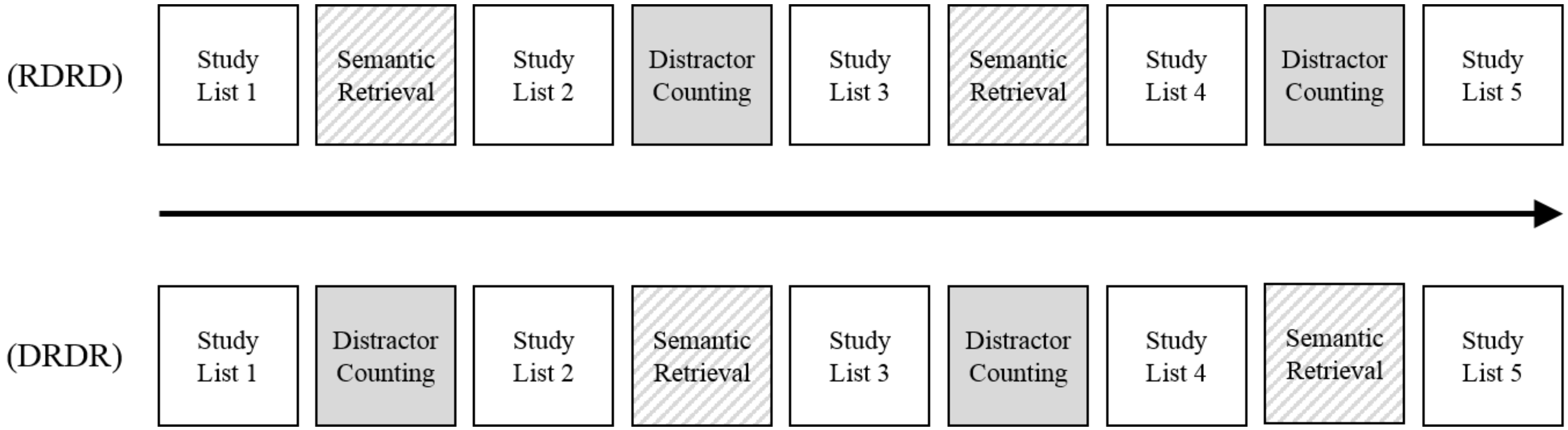


Figure 5. Study phase design in Experiment 2A by counterbalancing condition (RDRD or DRDR).

Figure 6

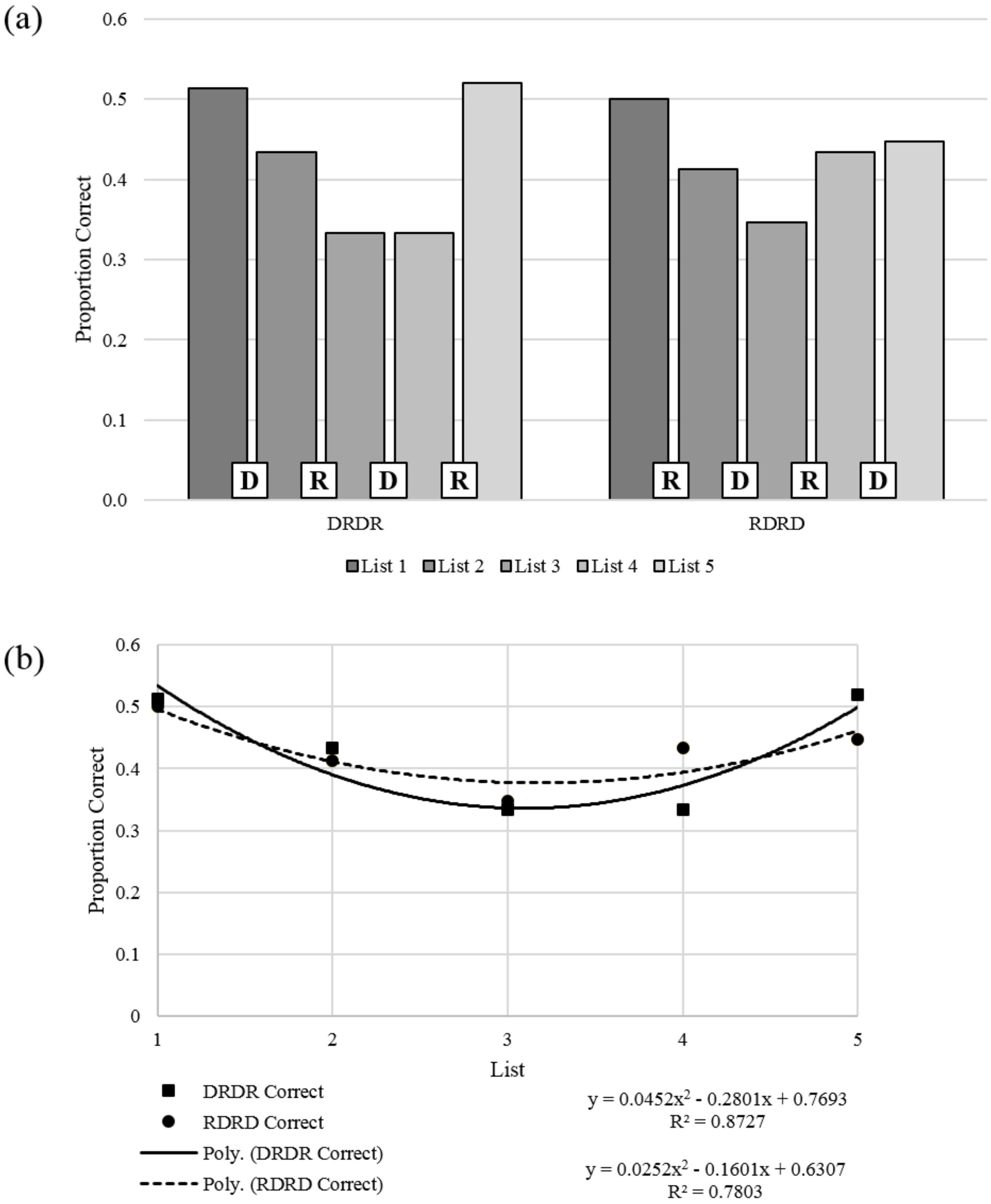


Figure 6. (a) Proportion correct as a function of list (1-5) and counterbalancing condition (RDRD or DRDR) for Experiment 2A. *(b)* Proportion correct as a function of list and condition in Experiment 2A fitted by quadratic regression line.

Figure 7

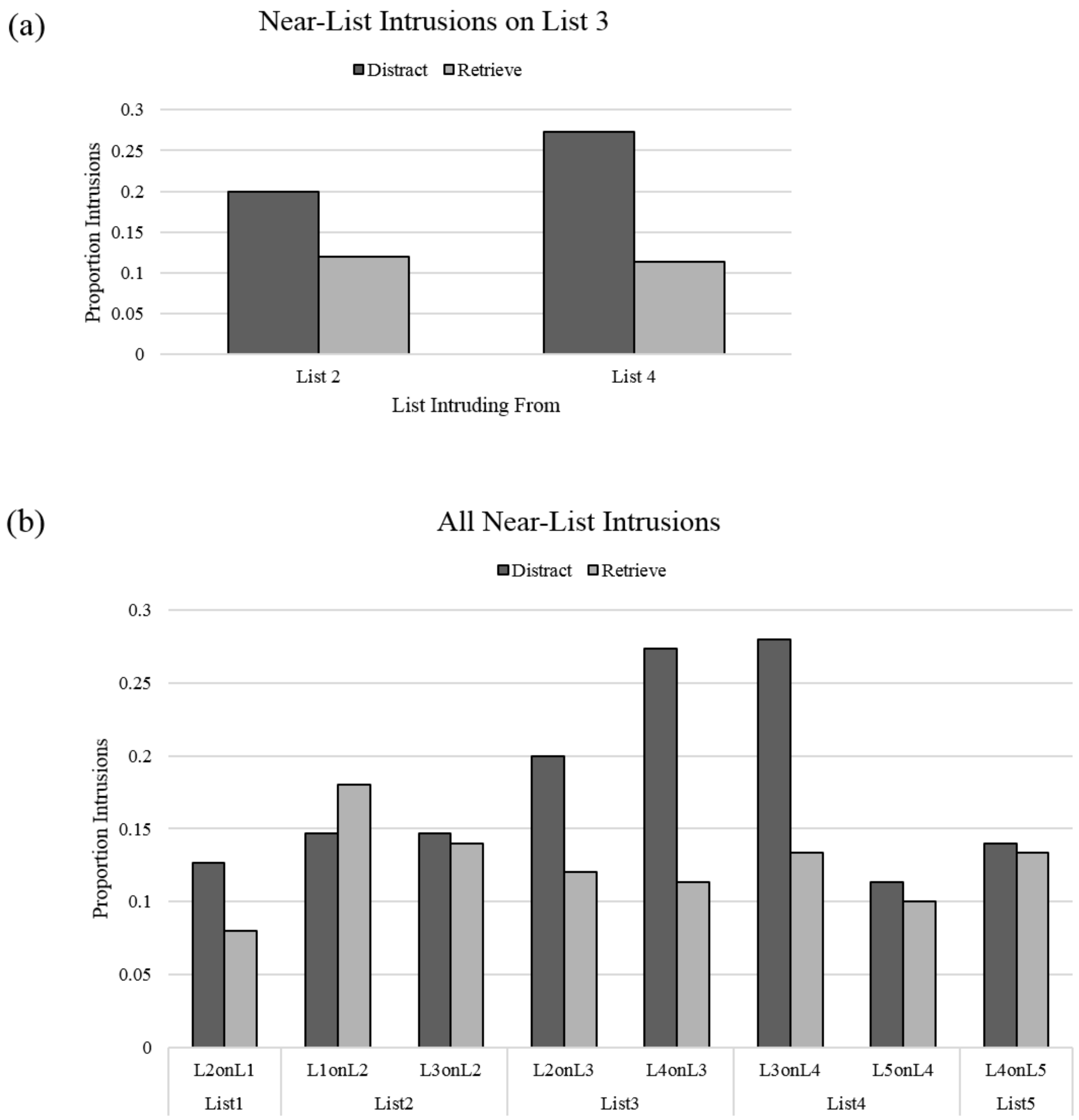

(a)
Near-List Intrusions on List 3
Distract
Retrieve
Proportion Intrusions
0.3
0.25
0.2
0.15
0.1
0.05
0
List 2
List 4
List Intruding From
(b)
All Near-List Intrusions
Distract
Retrieve
Proportion Intrusions
0.3
0.25
0.2
0.15
0.1
0.05
0
L2onL1
L1onL2
L3onL2
L2onL3
L4onL3
L3onL4
L5onL4
L4onL5
List1
List2
List3
List4
List5

Figure 7 (continued)

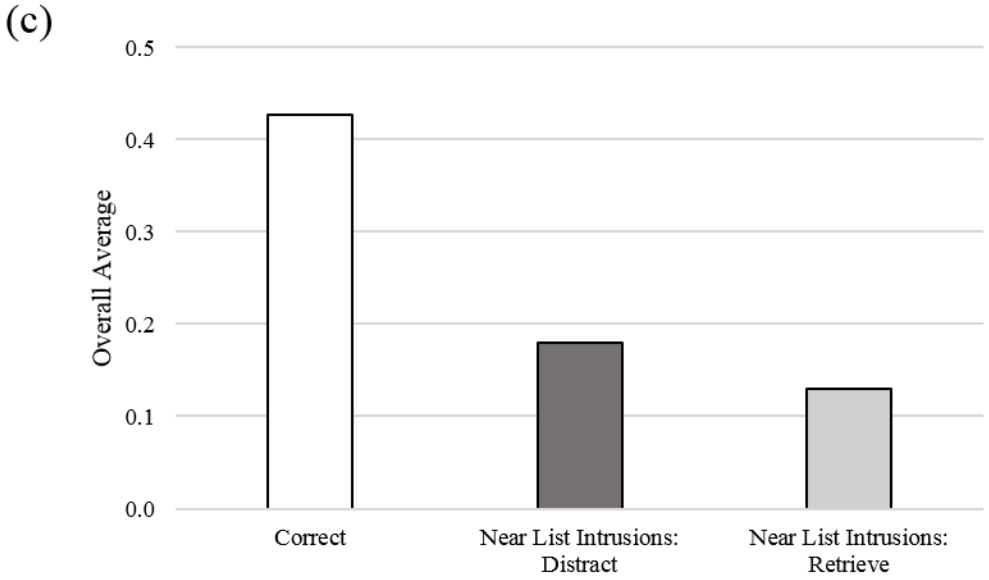


Figure 7. (a) Proportion of near-list intrusions onto List 3 as a function of source list (List 2 or List 4) and task (distract or retrieve) for Experiment 2A. *(b)* Proportion of near-list intrusions as a function of target list (1-5), source list (target list +/-1), and task (retrieve or distract) for Experiment 2A. The upper horizontal axis label represents the source of the intrusions (e.g., L2onL1 are List 2 items intruding on List 1); the lower horizontal axis label represents the target list. *(c)* Overall proportion correct and intrusions from near-lists (by intervening task) for Experiment 2A, collapsed across condition and list.

Figure 8

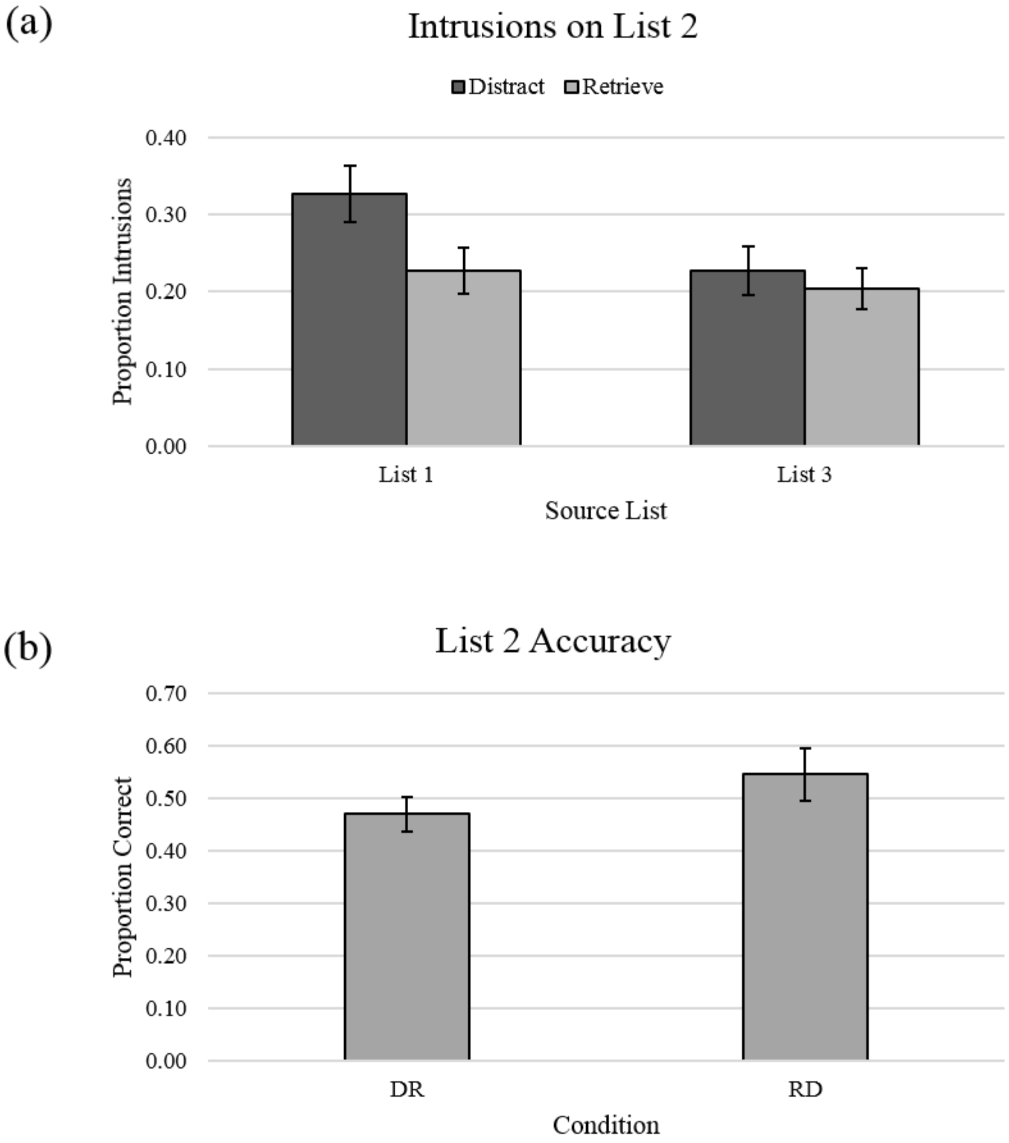


Figure 8. (a) Proportion of near-list intrusions onto List 2 as a function of source list (List 1 or List 3) and task (distract or retrieve) for Experiment 2B. *(b)* Proportion correct on List 2 as a function of condition (DR or RD) for Experiment 2B. Error bars represent standard error of the mean.

Figure 9

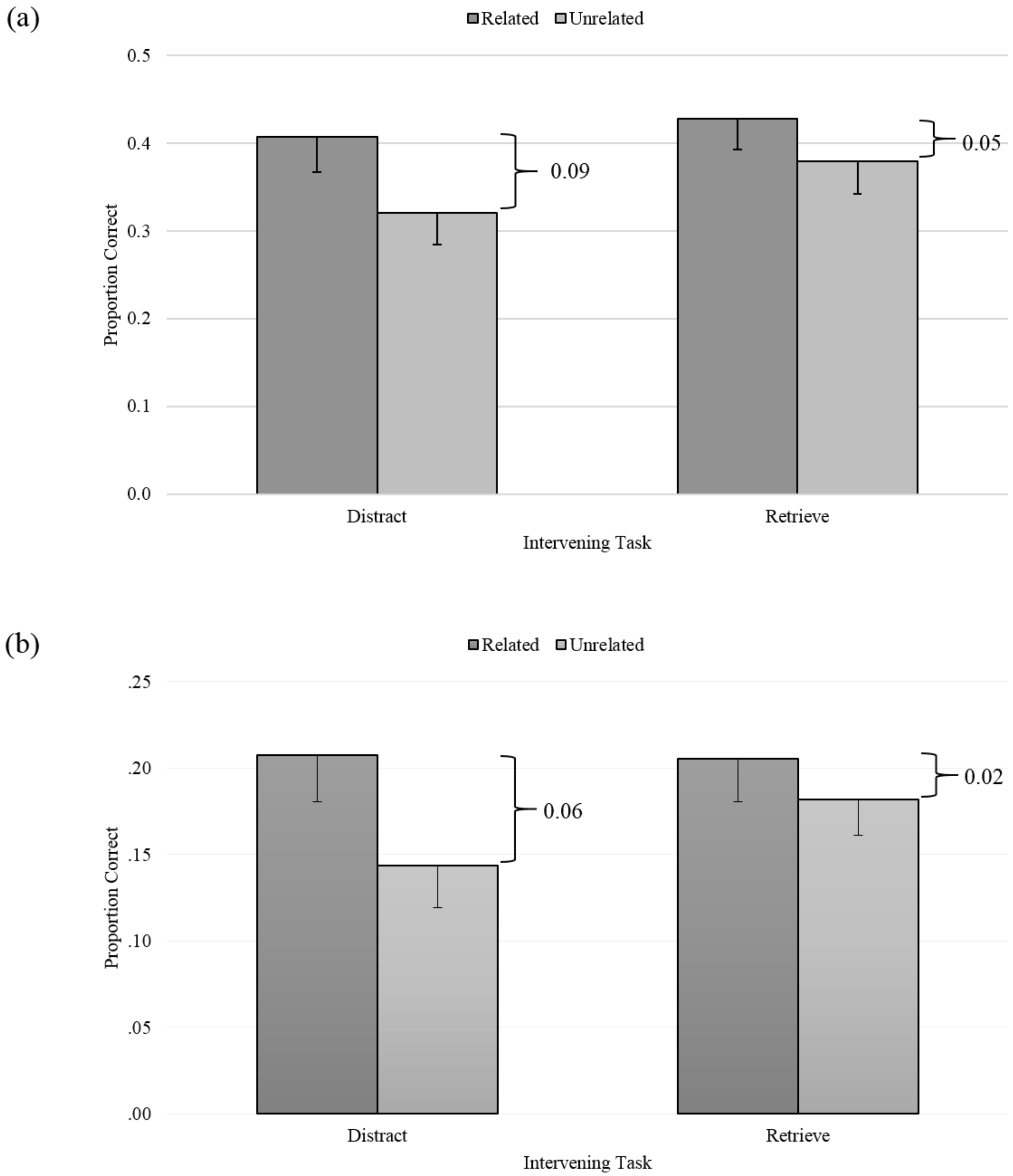


Figure 9. (a) Proportion correct on List 1 as a function of intervening task (distract or retrieve) and material type (related or unrelated) for the first pilot experiment examining related material. *(b)* Proportion correct on List 2 as a function of intervening task (distract or retrieve) and material type (related or unrelated) for the second pilot experiment examining related material. The difference between related and unrelated material is the reminding effect.

Figure 10

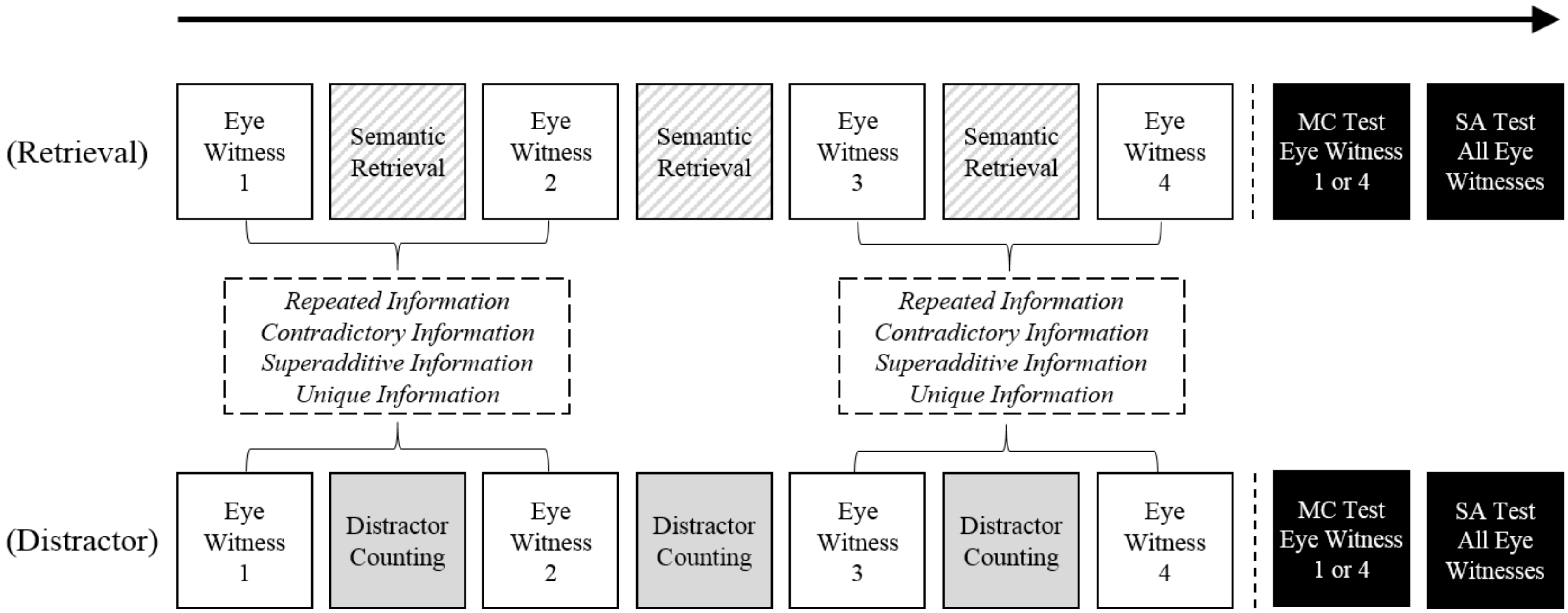


Figure 10. Experiment 3 design by condition (distract or retrieve). Repeated, contradictory, superadditive, and unique information was always paired between the first and second eye witnesses or the third and fourth eye witnesses. Participants were given a multiple choice test on either the first or last eye witness testimony followed by a short answer test on all of the eye witness testimonies.

Figure 11

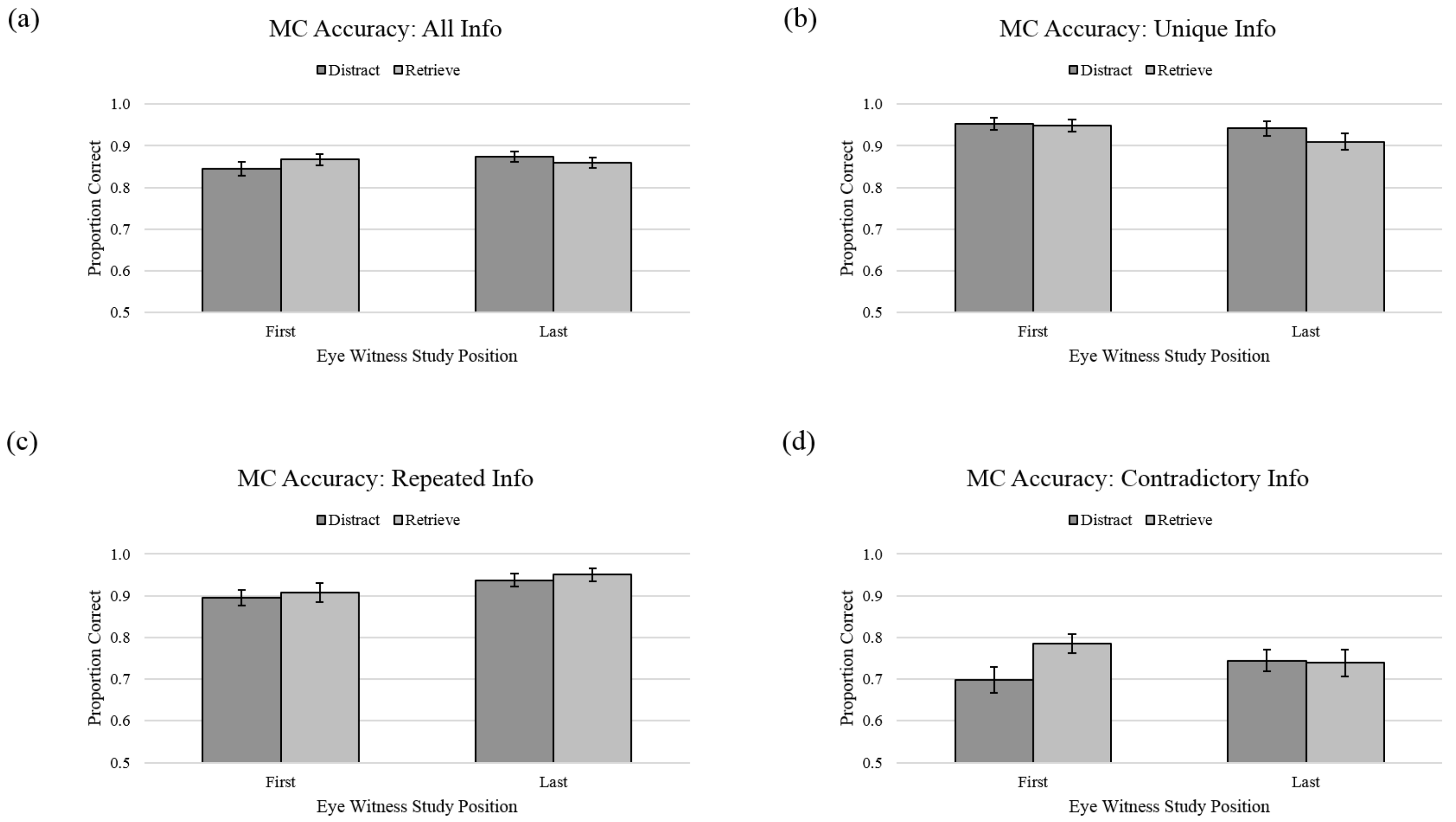


Figure 11. Accuracy on the multiple choice test in Experiment 3 by intervening task (distract or retrieve) for *(a)* all categories of information, *(b)* unique information only, *(c)* repeated information only, and *(d)* contradictory information only. Error bars represent standard error of the mean.

Figure 12

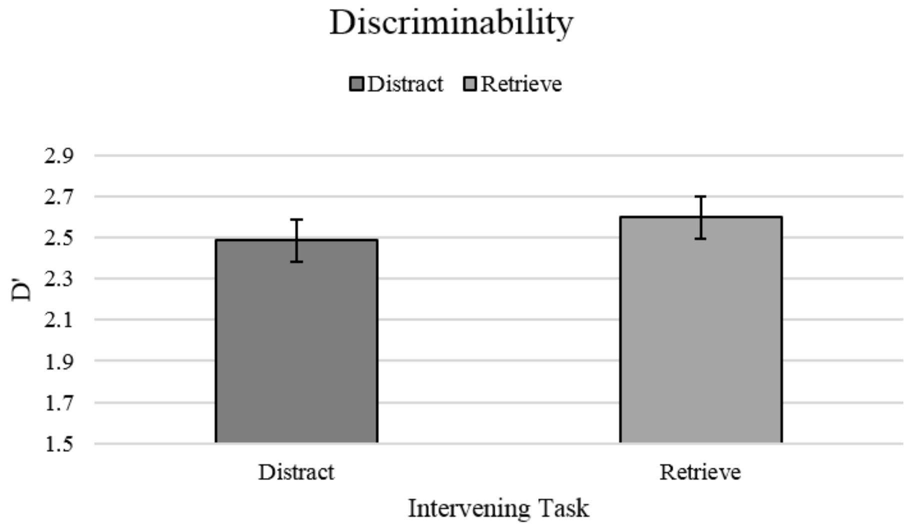


Figure 12. Discriminability on the individual portion of the matrix section of the short answer test in Experiment 3 by intervening task (distract or retrieve). Error bars represent standard error of the mean.

Figure 13

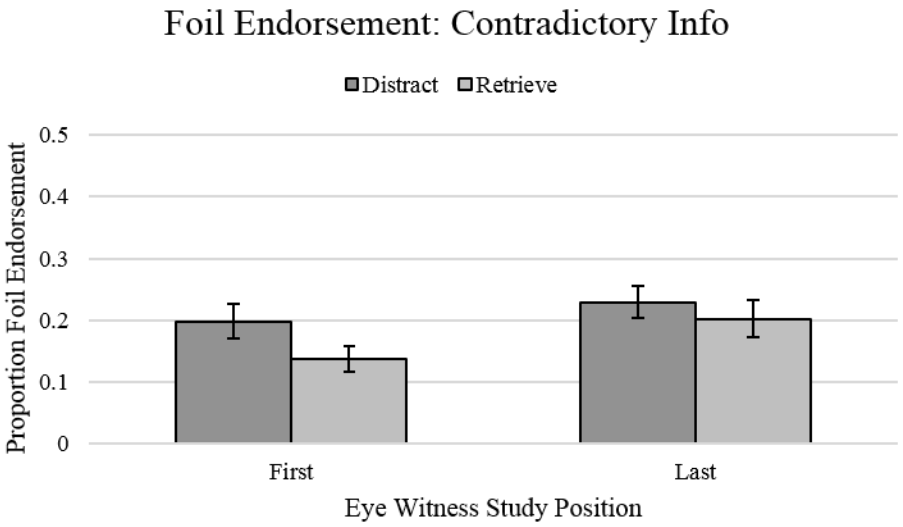


Figure 13. Endorsement of foils on multiple choice questions about contradictory information in Experiment 3 by intervening task (distract or retrieve). Error bars represent standard error of the mean.

Figure 14

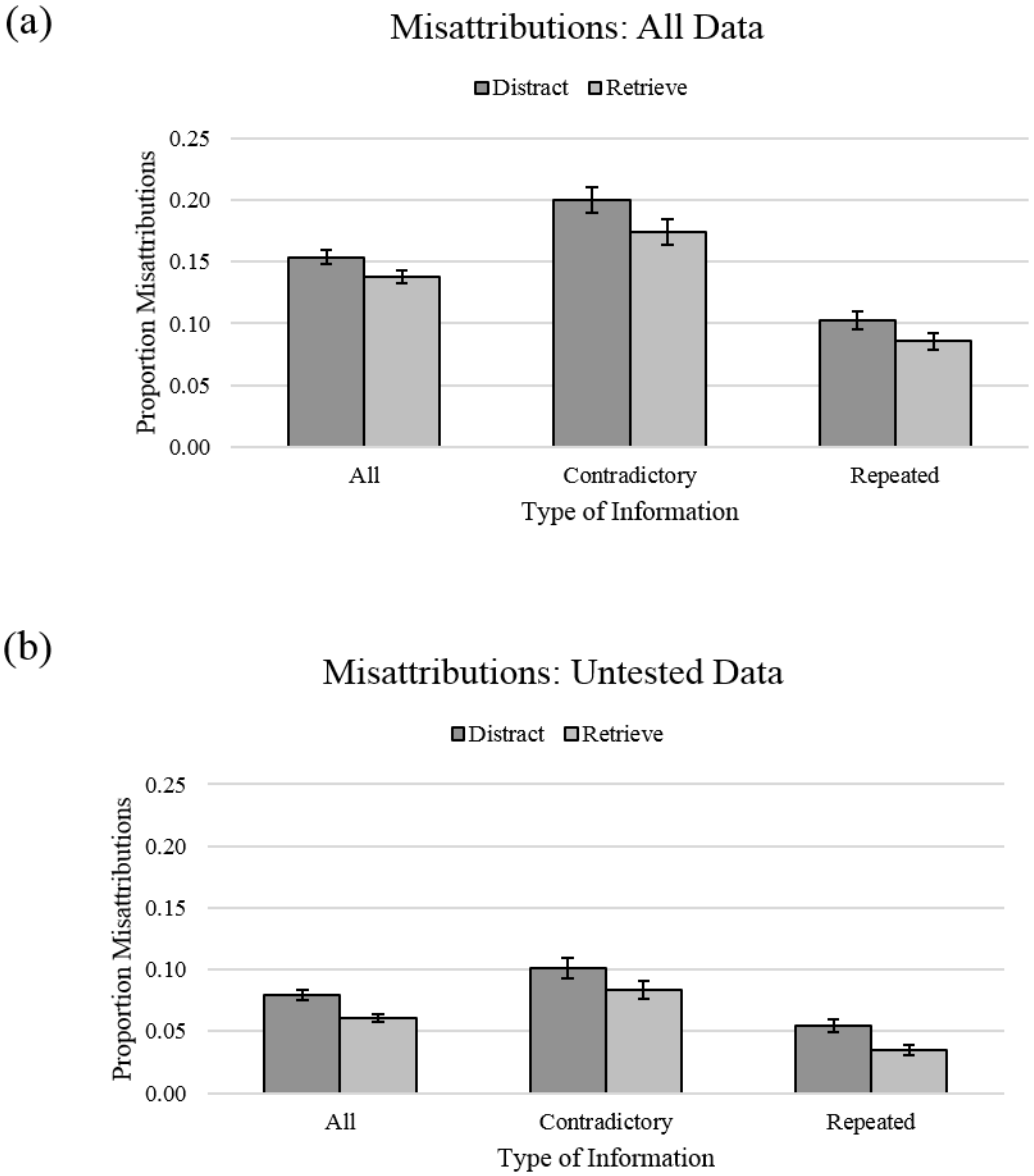


Figure 14. Misattributions on the individual portion of the matrix section of the short answer test in Experiment 3 by intervening task (distract or retrieve) for *(a)* all of the data and *(b)* only the data on information not already tested in the multiple choice test. Error bars represent standard error of the mean.

Figure 15

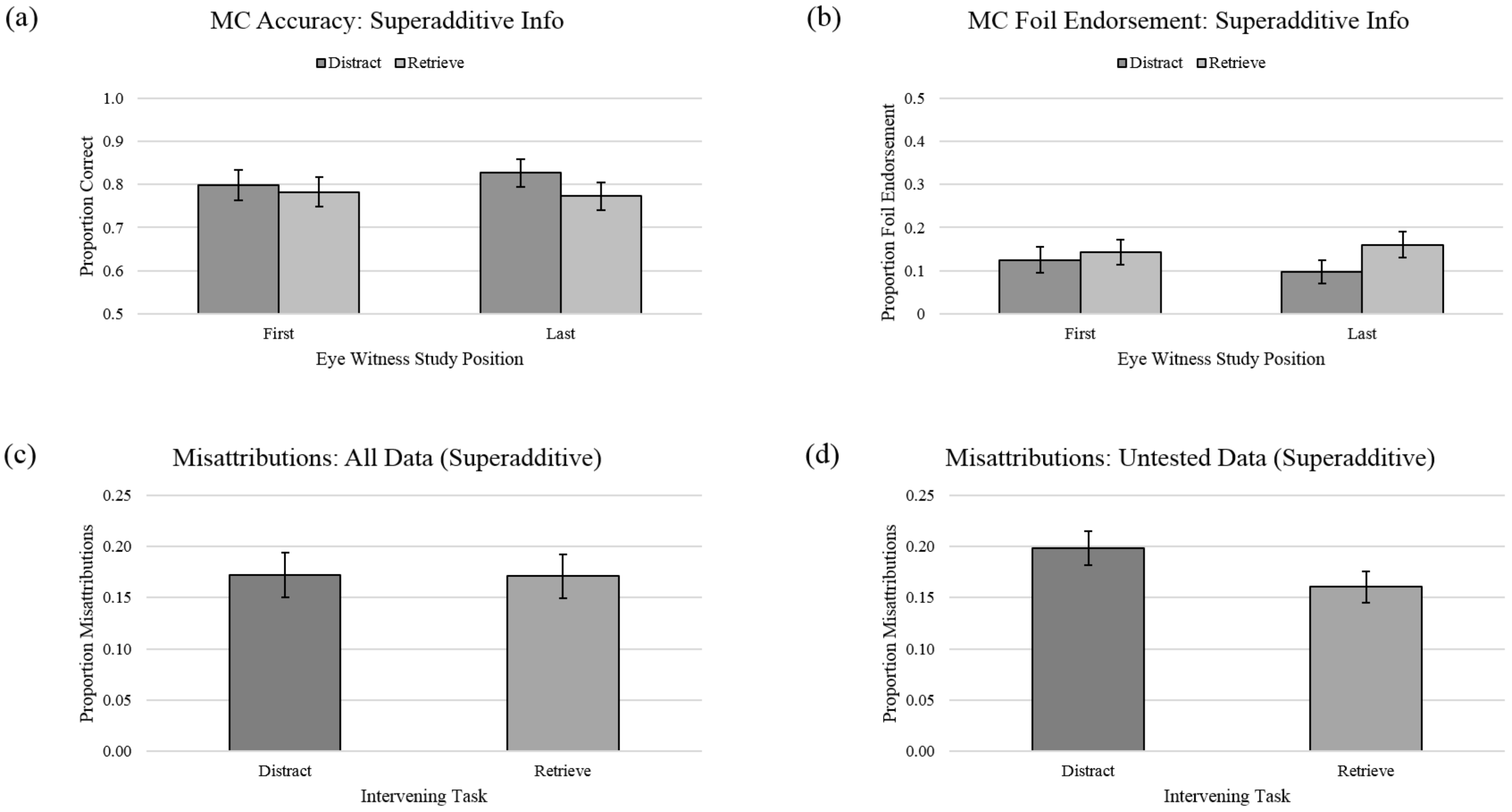


Figure 15. Results of analyses performed on superadditive information only for *(a)* accuracy on the multiple choice test by intervening task and eye witness study position, *(b)* foil endorsement in the multiple choice test by intervening task and study position, *(c)* misattributions for the short answer matrix test using all of the data by intervening task, and *(d)* misattributions for the short answer test using only information that was not also queried during the multiple choice test by intervening task. Error bars represent standard error of the mean.

Figure 16

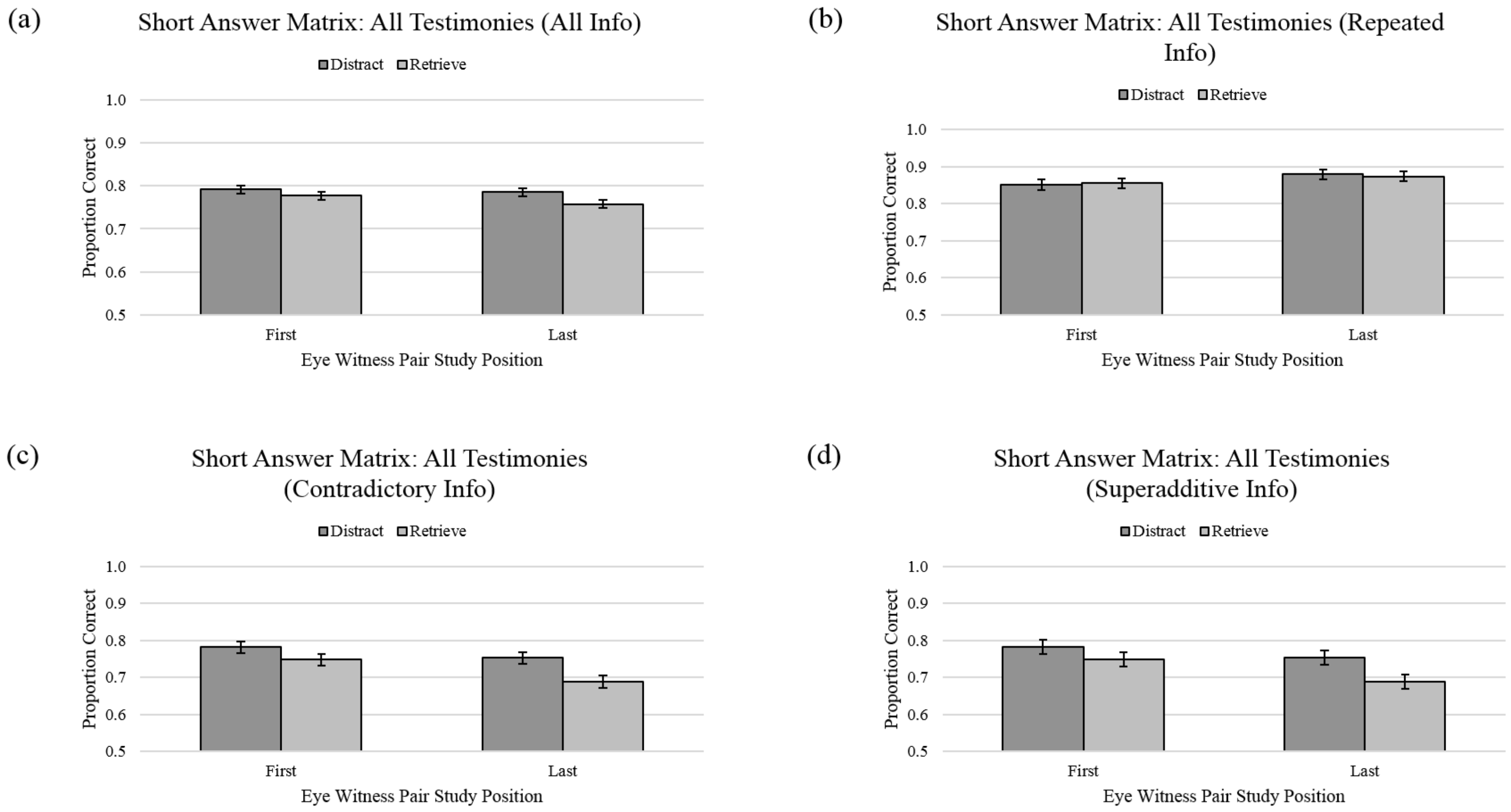


Figure 16. Performance on the short answer matrix test for the portion testing overall memory for the crime by intervening task (distract or retrieve) and study position of the eye witness pair (first or last) for *(a)* all types of information, *(b)* repeated information only, *(c)* contradictory information only, and *(d)* superadditive information only. Error bars represent standard error of the mean.

Figure 17

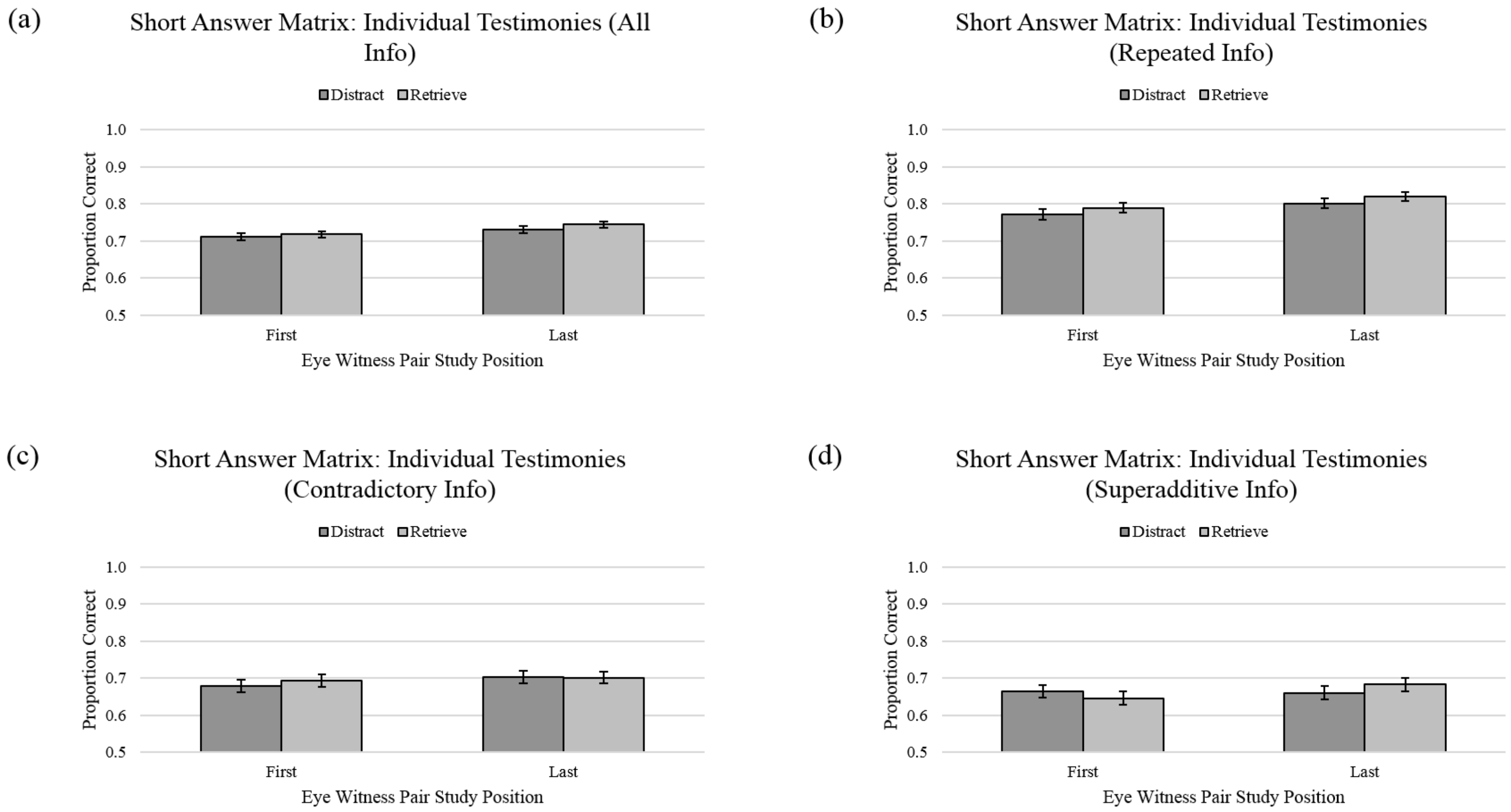


Figure 17. Performance on the short answer matrix test for the portion testing memory for individual eye witness testimonies by intervening task (distract or retrieve) and study position of the eye witness pair (first or last) for *(a)* all types of information, *(b)* repeated information only, *(c)* contradictory information only, and *(d)* superadditive information only. Error bars represent standard error of the mean.

Figure 18

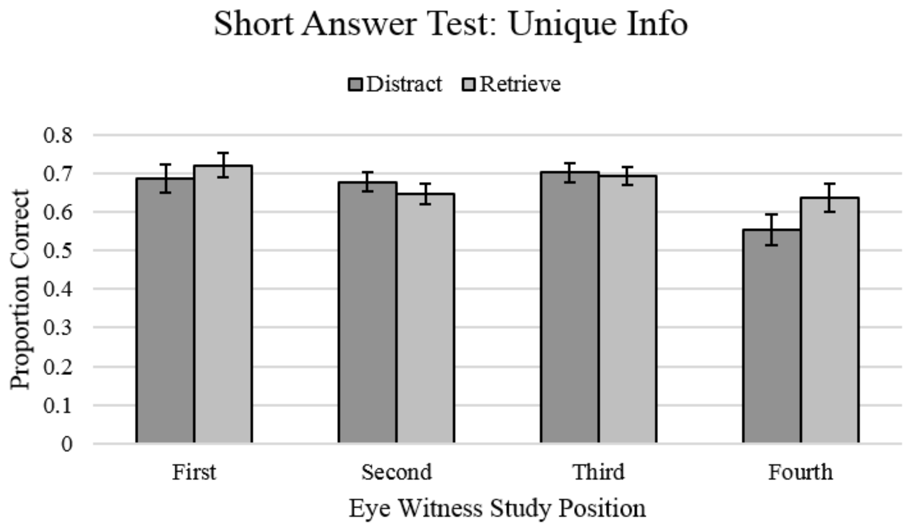


Figure 18. Performance on the short answer test on unique information by eye witness study position and intervening task (distract or retrieve). Error bars represent standard error of the mean.

Tables

Table 1

Analysis	Information Included	Level of Analysis	Results (After dropping first 10 subjects)
Accuracy (Multiple Choice Test)	All types of information	Simple effect for first position	*z* = 1.269, *p* = .204
		Simple effect for last position	*z* = .832, *p* = .405
	Unique information	Simple effect for first position	*z* < .001, *p* = 1.00
		Simple effect for last position	*z* = 1.58, *p* = .119
	Repeated information	Simple effect for first position	*z* = .392, *p* = .695
		Simple effect for last position	*z* = .606, *p* = .545
	Contradictory information	Simple effect for first position	*z* = 2.294, *p* = .022
		Simple effect for last position	*z* = .010, *p* = .992
Discriminability	All types of information	Overall	*t*(161.67) = .743, *p* = .459
Foil Endorsement (Multiple Choice Test)	Contradictory information	Overall	*z* = 1.632, *p* = .103
		Simple effect for first position	*z* = 1.629, *p* = .103
		Simple effect for last position	*z* = .638, *p* = .524
Misattributions (Short Answer Test: Individual	All types of information	Overall	*z* = 2.683, *p* = .007
	Contradictory information	Overall	*z* = 2.204, *p* = .028
	Repeated information	Overall	*z* = 1.949, *p* = .051
Misattributions (Short Answer Test: Individual	All types of information	Overall	*z* = 2.421, *p* = .016
	Contradictory information	Overall	*z* = 1.815, *p* = .070
	Repeated information	Overall	*z* = 1.995, *p* = .046

Table 1. Results of Experiment 3 analyses after dropping the first ten (pilot) subjects.

Table 2

Bayes Factor (*B_{10}*)	**Evidence against *H_0***
1 to 3	Weak/Anecdotal
3 to 20	Positive/Substantial
20 to 150	Strong
> 150	Very strong

Table 2. Interpretation of Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995).

Table 3

Analysis	List	Conditions	Mixed Effects Model	Student's T-Test	B_{10} (based on t-test)	Interpretation of B_{10}
Proportion Correct	List 1	Pure Distract vs. Pure Retrieve	*z* = 2.120, *p* = .034	*t*(31.33) = 1.5392, *p* = .134	0.811	Weak for H_0
		Pure Distract vs. Switch	*z* = 3.569, *p* < .001	*t*(30.13) = 2.842, *p* = .008	6.79	Substantial for H_a
		Pure Retrieve vs. Switch	*z* = 1.092, *p* = .275	*t*(26.65) = .7857, *p* = .439	0.377	Weak for H_0
	List 5	Pure Distract vs. Pure Retrieve	*z* = 5.672, *p* < .001	*t*(25.33) = 3.253, *p* = .003	13.975	Substantial for H_a
		Pure Distract vs. Switch	*z* = 4.750, *p* < .001	*t*(39.49) = 3.444, *p* = .001	26.829	Strong for H_a
		Pure Retrieve vs. Switch	*z* = 1.803, *p* = .073	*t*(24.47) = 1.001, *p* = .327	0.44	Weak for H_0
		Retrieve-Distract vs. Distract-Retrieve	*z* = 3.018, *p* = .003	*t*(31.86) = 2.007, *p* = .053	0.129	Substantial for H_0
Interlist Intrusions	List 1	Pure Distract vs. Pure Retrieve	*t*(136) = .275, *p* = .784	*t*(120.85) = .592, *p* = .555	0.377	Weak for H_0
		Pure Distract vs. Switch	*t*(136) = .396, *p* = .692	*t*(97.46) = .822, *p* = .413	0.386	Weak for H_0
		Pure Retrieve vs. Switch	*t*(136) = .079, *p* = .937	*t*(126.46) = .213, *p* = .831	0.299	Substantial for H_0
	List 5	Pure Distract vs. Pure Retrieve	*t*(136) = 2.884, *p* = .005	*t*(96.53) = 2.140, *p* = .035	1.819	Weak for H_a
		Pure Distract vs. Switch	*t*(136) = 3.331, *p* = .001	t(80.11) = 2.269, *p* = .026	2.245	Weak for H_a
		Pure Retrieve vs. Switch	*t*(136) < .001, *p* = 1.00	*t*(123.34) < .001, *p* = 1.00	0.294	Substantial for H_0

Table 3. Statistics based on mixed effects model and traditional t-tests, along with Bayes factors statistic and interpretation for Experiment 1.

Table 4

Analysis	**Level of Analysis**	**Mixed Effects Model**	**Student's T-Test**	$\boldsymbol{B_{10}}$ **(based on t-test)**	**Interpretation of** $\boldsymbol{B_{10}}$
Near-list intrusions onto List 3	Overall	$t(30) = 3.006$, $p = .005$	$t(29) = 2.921$, $p = .007$	6.343	Substantial for H_a
	Simple effect for List 2	$t(59.96) =1.400$, $p = .167$	$t(23.01) = 1.702$, $p = .102$	0.875	Weak for H_0
	Simple effect for List 4	$t(59.96) = 2.800$, $p = .007$	$t(16.96) = 2.311$, $p = .034$	2.353	Weak for H_a
All near-list intrusions	Overall	$t(240) = 2.602$, $p = .010$	$t(195.04) = 2.591$, $p = .010$	3.279	Substantial for H_a
Near-list intrusions onto List 3 (including performance on source	Overall	$t(29.28) = 2.941$, $p = .006$	$t(29) = 2.921$, $p = .007$	6.343	Substantial for H_a
	Simple effect for List 2	$t(47.95) = 2.068$, $p = .044$	$t(23.01) = 1.702$, $p = .102$	0.875	Weak for H_0
	Simple effect for List 4	$t(49.64) = 2.195$, $p = .033$	$t(16.96) = 2.311$, $p = .034$	2.353	Weak for H_a
Accuracy on List 1	Overall	$z = .082$, $p = .934$	$t(27.51) = .121$, $p = .904$	0.346	Weak for H_0

Table 4. Statistics based on mixed effects model and traditional t-tests, along with Bayes factors statistic and interpretation for Experiment 2A.

Table 5

Analysis	**Level of Analysis**	**Mixed Effects Model**	**Student's T-Test**	**B_{10} (based on t-test)**	**Interpretation of B_{10}**
Interlist intrusions onto List 2	Overall	$t(80) = 2.017, p = .047$	$t(39) = 1.867, p = .069$	0.824	Weak for H_0
	Simple effect for List 1	$t(80) = 2.313, p = .023$	$t(36.20) = 2.107, p = .042$	1.714	Weak for H_a
	Simple effect for List 3	$t(80) = .540, p = .591$	$t(36.81) = .569, p = .573$	0.351	Weak for H_0
Accuracy on List 2	Overall	$z = 1.351, p = .177$	$t(32.95) = 1.288, p = .207$	0.594	Weak for H_0

Table 5. Statistics based on mixed effects model and traditional t-tests, along with Bayes factors statistic and interpretation for Experiment 2B.

Table 6

Analysis	Information Included	Level of Analysis	Mixed Effects Model	Student's T-Test	***B*** $_{10}$ **(based on t-test)**	**Interpretation of *B*** $_{10}$
Accuracy (Multiple Choice Test)	All types of information	Simple effect for first position	*z* = 1.105, *p* = .269	*t*(93.17) = .920, *p* = .329	0.328	Substantial for H_0
		Simple effect for last position	*z* = .801, *p* = .423	*t*(93.00) = .822, *p* = .413	0.291	Substantial for H_0
	Unique information	Simple effect for first position	*z* = .188, *p* = .851	*t*(94.99) = .204, *p* = .839	0.218	Substantial for H_0
		Simple effect for last position	*z* = 1.270, *p* = .204	*t*(88.09) = 1.346, *p* = .182	0.479	Weak for H_0
	Repeated information	Simple effect for first position	*z* = .565, *p* = .572	*t*(92.91) = .421, *p* = .675	0.231	Substantial for H_0
		Simple effect for last position	*z* = .638, *p* = .524	*t*(93.00) = .596, *p* = .552	0.252	Substantial for H_0
	Contradictory information	Simple effect for first position	*z* = 2.253, *p* = .024	*t*(86.46) = 2.247, *p* = .027	1.942	Weak for H_A
		Simple effect for last position	*z* = .147, *p* = .896	*t*(89.04) = .131, *p* = .896	0.217	Substantial for H_0
Discriminability	All types of information	Overall	*N/A*	*t*(170.88) = .775, *p* = .440	0.217	Substantial for H_0
Foil Endorsement (Multiple Choice Test)	Contradictory information	Overall	*z* = 1.857, *p* = .063	*t*(189.84) = 1.683, *p* = .094	0.586	Weak for H_0
		Simple effect for first position	*z* = 1.861, *p* = .062	*t*(87.27) = 1.735, *p* = .086	0.804	Weak for H_0
		Simple effect for last position	*z* = .726, *p* = .468	*t*(90.39) = .685, *p* = .495	0.265	Substantial for H_0
Misattributions (Short Answer Test: Individual Matrix Section): All Data	All types of information	Overall	*z* = 2.318, *p* = .021	*t*(170.13) = 1.16, *p* = .247	0.307	Substantial for H_0
	Contradictory information	Overall	*z* = 1.994, *p* = .046	*t*(170.77) = 1.312, *p* = .191	0.365	Weak for H_0
	Repeated information	Overall	*z* = 1.804, *p* = .071	*t*(170.91) = 1.15, *p* = .521	0.304	Substantial for H_0
Misattributions (Short Answer Test: Individual Matrix Section): Only Untested Pairs	All types of information	Overall	*z* = 2.499, *p* = .012	*t*(170.93) = 1.92, *p* = .056	0.906	Weak for H_0
	Contradictory information	Overall	*z* = 1.460, *p* = .144	*t*(164.06) = .913, *p* = .363	0.242	Substantial for H_0
	Repeated information	Overall	*z* = 2.289, *p* = .022	*t*(170.62) = 1.934, p = .055	0.924	Weak for H_0

Table 6. Statistics based on mixed effects model and traditional t-tests, along with Bayes factors statistic and interpretation for Experiment 3.