DISSERTATION

ITEM CONTENT VERSUS CONTEXTUAL STRENGTHENING FOLLOWING RETRIEVAL

Submitted by

Christopher A. Rowland

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2015

Doctoral Committee:

Advisor: Edward DeLosh

Matthew Rhodes
Bryan Dik
Daniel Robinson

ABSTRACT

ITEM CONTENT VERSUS CONTEXTUAL STRENGTHENING FOLLOWING RETRIEVAL

Despite a substantial literature describing the memory benefit resulting from testing (i.e., memory retrieval), relatively few investigations have attempted to detail how retrieval acts as a memory modifier. One core issue concerns the extent to which testing and studying effect fundamentally similar or different processes or components of memories. The present paper introduces two computational models, both based in REM theory (Shiffrin & Steyvers, 1997) and designed to provide a plausible basis for describing the testing effect at a more mechanistic level than existing theories. The two models are derived from the same set of core assumptions about the functioning of the memory system, and differ only in their specifications of the components of memories that are modified as a result of retrieval. The “Item Model” (IM) assumes that retrieval serves primarily to strengthen the target item content representation of information that is retrieved. In contrast, the “Context Model” (CM) assumes that retrieval serves to embed additional contextual information into the target memory trace, facilitating the subsequent ability of the memory system to locate such items. This manuscript provides coverage of relevant areas in the literature that have bearing on the IM and CM, details the implementation of the models and their larger framework, and reports on 4 experiments designed to test contrasting predictions of the IM and CM. Experiment 1 observed a testing effect using a mixed list, but not a pure list design, implying that testing may serve to enhance the search process by strengthening context information in memory. Experiments 2-4 were designed to examine the effects of reinstating contextual information during final testing on the testing effect.

Experiments 2 and 3 found that reinstating either perceptual contextual elements (Exp. 2), or semantic context cues (Exp. 3) at the time of final test did not significantly impact the magnitude of the testing effect. However, Experiment 4 found that reinstating the initial learning mental/temporal context at the time of final test mitigated the magnitude of the testing effect. Potential nuanced interactions between testing and context in memory are discussed.

TABLE OF CONTENTS

INTRODUCTION

Retrieval acts as a potent memory modifier. The act of remembering a past event can serve, in many cases, to strengthen one's retention of the remembered event such that it will be more likely recalled in the future. This positive effect of retrieval on subsequent memory is referred to as the *testing effect* (see reviews from Roediger & Karpicke, 2006a, and Rowland, 2014), research upon which has strongly intensified in the past decade (Rawson & Dunlosky, 2011). Although much is known about the circumstances under which testing is more or less beneficial to memory, less attention has been called to the question of *how* testing strengthens memory.

The notion of “strengthening” a memory (through testing or other means) is its self an ambiguous descriptor (Tulving & Bower, 1974). Typically, an operation can be considered to strengthen a memory if it serves to increase the likelihood of accessing that memory at some point in the future. However, there are a variety of ways in which a memory may be strengthened. For example, a memory trace may receive additional content in its representation; more effective cues may be present to guide retrieval, or interference from other competing memories may be reduced, all of which serve to increase the likelihood of remembering. Furthermore, in the view of some frameworks, increasing the strength of one aspect of a memory may not immediately manifest in increased recall likelihood (e.g., see the distinction between storage and retrieval strength, Bjork & Bjork, 1992). Existing work on the testing effect is largely agnostic as to how memory strengthening occurs, and what precisely is strengthened as a result of retrieval.

Most examinations of the testing effect can be described as using a common experimental paradigm. Participants begin by being exposed to a set of material that they are tasked with learning, such as a list of words or word pairs (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2010; Rowland & DeLosh, 2014a; Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014; Toppino & Cohen, 2009), essays or more complex prose passages (e.g., Roediger & Karpicke, 2006b), or non-verbal stimuli such as maps or images (e.g., Kang, 2010; Rohrer, Taylor, & Sholar, 2010). Following this initial study phase, and typically a delay, participants are re-exposed to the material either by being granted a second study opportunity (i.e., a "restudy" condition), or are given an initial test over the materials (i.e., a "testing" condition). Subsequently, after an additional delay (i.e., a retention interval), participants' memory is assessed during a final test (i.e., a memory assessment). Thus, empirically, the testing effect refers to the reliable finding that participants in the testing condition outperform participation in the restudy condition on the final memory assessment. That is, testing "strengthens" memory, in some way.

Most of the theoretical accounts that have been advanced to describe the testing effect have been proposed at a broad level, lacking mechanistic specificity. For instance, testing during learning has been proposed to induce a similarity in the type of processing during learning that is then subsequently utilized during a final memory test (see Bjork, 1988; Roediger & Karpicke, 2006a). Other accounts of the testing effect assume that the benefit to memory results from, or is related to, the more effortful or difficult processing that a retrieval task demands, when compared with a restudy task (e.g., Bjork, 1978; Jacoby, 1978; Pyc & Rawson, 2009). Although such types of theories can serve as useful characterizations of the testing effect (see Rowland, 2014, for a more thorough treatment of such theories), the common thread amongst most existing theoretical work is that the proposed components driving the testing effect are characterized at only a broad, abstracted level.

There have been only a limited number of theoretical accounts of the testing effect that attempt to describe the ways in which retrieval impacts memory at a fine-grained process or mechanistic level, and in most cases, such accounts assume that testing serves to alter the way that target items are themselves represented in memory. One such account, the mediator effectiveness hypothesis (Carpenter, 2011; Pyc & Rawson, 2010; see also the more general elaborative retrieval hypothesis, Carpenter, 2009; Carpenter & DeLosh, 2006), is rooted in the notion that retrieval exploits the semantic characteristics of information one is attempting to learn. In particular, the mediator effectiveness hypothesis specifies that testing during learning can facilitate a learners' ability to generate "mediating information," defined as information (e.g., a word or concept) that forms a link between a cue and a target piece of information. For example, given a cue-target pair of "CAT – BOWL," an initial test during learning given the cue "CAT - ?" may prompt the learner to generate a mediator linking the cue and target together, such as "food." Later, at final test, given "CAT – ?," the learner may recall the mediator "food," thereby increasing the likelihood of recalling the target "BOWL." In broad terms, the mediator effectiveness hypothesis and related "elaborative retrieval" accounts (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006; Rawson, Vaughn, & Carpenter, 2014) assume that testing during learning exploits the semantic characteristics of material that is being learned, such that the content of memories is embellished, strengthened, or elaborated upon.

A similar characterization of the testing effect that is also rooted in the utilization of semantic information comes from Verkoeijen, Bouwmeester, and Camp (2012; see also, Bouwmeester & Verkoeijen, 2011). Their account assumes that retrieval serves to strengthen the representation of semantic features in a memory trace, whereas restudy strengthens surface features (e.g., perceptual characteristics). As such, when a final test is administered that allows for the exploitation of surface-level cues to guide retrieval, restudy may emerge as a more effective learning strategy. In contrast, given conditions in which participants must utilize semantics to guide retrieval, a testing effect should emerge. In a novel test of this theory, Verkoeijen et al. (2012) had bilingual participants (fluent in Dutch and English) learn semantically themed word lists in Dutch through study and either initial testing or restudying. After a brief delay, participants were given a recognition test over the stimuli in either Dutch (i.e., the same language as learning) or English (i.e., different language from learning). Verkoeijen et al. (2012) observed a testing effect only for participants given a different language test, presumably because phonological and orthographic cues were no longer present in the recognition test stimuli (given that the words were presented in a different language). Again, the theory assumes that testing exploits the content of items stored in memory: in this case, their semantic content.

Theoretical characterizations that emphasize the role of semantics in the testing effect have drawn substantial attention and support. In addition to the characterizations outlined above, additional findings from the literature have been advanced in the same vein. For instance, testing may facilitate the semantic organization of information in memory (Congleton & Rajaram, 2011, 2012; Zaromb & Roediger, 2010); promote false recall of information that is semantically related to learned material (McDermott, 2006, though cf. Nunes & Weinstein, 2012); and activate semantically related material that was learned during the same episode (e.g., Chan, McDermott, & Roediger, 2006; Little, Storm, & Bjork, 2011). As such, the emphasis of theoretically motivated research concerning the testing effect has, largely, been to assume a central role for semantics, or more generally, target item characteristics, in driving the testing effect.

Despite the strengths of semantic-based accounts, one difficulty faced by characterizations of the testing effect that emphasize the exploitation of semantic item characteristics involves the emergence of testing effects in situations devoid of semantic content. Reliable testing effects emerge in studies where the material learned includes faces or names (e.g., Carpenter & DeLosh, 2005; Morris, Fritz, Jackson, Nichol, & Roberts, 2005; Sensenig, Littrell-Baez, & DeLosh, 2011), unfamiliar symbols e.g., (Kang, 2010), or spatial relationships (Carpenter & Kelley, 2012). In such cases, there is not clear, inherent, semantic content embedded in the material to be learned, yet a testing effect reliably emerges in such cases, and is not substantially different in magnitude compared with circumstances in which materials allowing for semantic exploitation are utilized (Rowland, 2014). As such, although testing may utilize semantic information embedded in learned material, it alone does not appear able to explain the emergence of the different types of testing effects observed in the literature.

A recent theoretical framework, applied to the testing effect, provides further illustration of the emphasis of retrieval influencing item-level characteristics. The bifurcation framework (Kornell, Bjork, & Garcia 2011; Halamish & Bjork, 2011) represents material to be learned in terms of tested and restudied item distributions across a continuum of memory strength. The framework assumes that, during learning, all items represented by the restudy item distribution receive an increment to memory strength by virtue of re-presentation (i.e., the entire restudy distribution is positively translated on the memory strength continuum). In contrast, the test item distribution becomes bifurcated (i.e., split) following the initial testing phase during learning. The proportion of the test item distribution that lies above an initial test strength threshold (i.e., the items that are successfully remembered on the initial test) are granted a large increment to their memory strength, whereas the portion of the test item distribution below threshold (i.e., not successful recalled) receives no strengthening from the initial test. Thus, the test item distribution becomes split into two, representing successfully, and unsuccessfully, retrieved items. At final test, the items in either the test or restudy item distributions with memory strengths that fall above a final test threshold are successfully recalled, whereas items below the strength threshold necessary for final test recovery are not recalled. The bifurcation framework thus characterizes item distributions in terms of their memory strength, and has been successful at explaining a number of interactions reported in the literature in a parsimonious way (see Halamish & Bjork, 2011; Kornell et al., 2011; Rowland, 2014; Rowland & DeLosh, 2014a). One strength of the framework lies in the fact that it assumes only that items themselves are strengthen by testing, to some degree beyond restudy, but otherwise are treated in a qualitatively similar way to restudy items. As such, the bifurcation model serves as another characterization of the testing effect that assumes that retrieval serves to exploit, alter, or strengthen item characteristics (although it is neutral as to the specific contributing mechanisms). Together, the theoretical characterizations described above share the commonality of assuming that testing impacts or exploits the content of items stored in memory.

The Importance of Context

Despite the emphasis in the existing testing effect literature that grants explanatory power of the effect to the exploitation of semantic or other item characteristics, there is evidence that the testing effect may emerge in part through enhancing the utility of episodic, contextual information that is linked to a target item. Generally, context is seen as having a central, fundamental role in learning and memory. The concept is rooted in the fact that learning and memory does not occur in a void, but rather material is exposed- and memory is utilized- in a physical, mental, and temporal context. Although a stringent, definitive explanation of what precisely entails "context" is lacking throughout the literature, it is typically viewed as at least encompassing elements that co-occur with the material that sits at the locus of attention. Context, in this sense, may include environmental stimuli peripheral to the attended material (e.g., the physical location one is in during learning; see, e.g., Smith, 1979); internal mood states or thoughts (e.g., a participants' stream of consciousness; see, e.g., Bower, 1972), and the given moment in time (i.e., temporal context, see, e.g., Howard & Kahana, 2002). Thus, in all cases, context refers to elements that co-occur with to-be-learned items (i.e., items at the focus of attention). Most modern episodic memory models assume that material is encoded into memory along with contextual elements that co-occur during learning, and that memories are retrieved by utilizing context as a potential cue to access past information, as will be elaborated upon when describing the models below. Context, in this sense, plays a central role in the functioning of episodic memory.

There is a variety of evidence that retrieval from episodic memory serves to influence our subsequent access to- and utilization of- context when probing memory. Initial testing seems to enhance the utility of recollection on subsequent memory decisions (Chan & McDermott, 2007). Recollection, in this context, refers to a memory process which allows one to remember details that co-occurred with the original encoding episode of a remembered item (e.g., context). Across a series of experiments, Chan and McDermott (2007) found that initial testing increased the contribution of recollection (see Jacoby, 1991) in a subsequent recognition task (see also Verkoeijen, Tabbers, & Verhange, 2011); yielded a greater likelihood of participants self-reporting that they remembered contextual details from the original study event; and enhanced list-differentiation. That is, initial testing facilitated the ability of participants to discriminate the original list context in which a given item was presented in, given a situation where stimuli were presented in multiple, separable, lists (see also Brewer, Marsh, Meeks, & Clark-Foos, 2010). Such findings suggest that testing may operate not solely through exploiting item-specific characteristics of tested information, but through altering the means and precision with which information is located in memory. That is, retrieval may help one subsequently hone in on the context in which a memory was previously encoded, thus improving memory performance by recreating a past context, or narrowing the "search set" (see Raaijmakers & Shiffrin, 1981): the set of potential contenders from which the memory system must discriminate between when searching for a target memory.

Additional support for this possibility comes from studies that have examined the effects of retrieval on interference in memory. Szpunar, McDermott, and Roediger (2008) developed a paradigm in which participants study a series of five lists. One group of participants studies each of the five lists in succession, whereas a different group is given an initial test over each list immediately after it is presented. After the five lists are presented, a final, criterial test is administered, in which participants are asked to recall only the items from list 5 (i.e., the final list). Work using this design shows that initial tests protect against the build-up of proactive interference, such that very few intrusions from lists 1-4 emerge on the criterial list 5 test for participants in the testing condition, relative to the substantial intrusions that emerge for participants not given the initial tests (Szpunar et al., 2008; see also, Weinstein, McDermott, & Szpunar, 2011). As such, testing appears to help participants discriminate between items that belong to given contexts (e.g., list contexts) from each other, and in doing so, allows for a narrowing of the possible candidate items to recall given a constrained retrieval task (see also Potts & Shanks, 2012).

The evidence reviewed above suggests that testing may serve to exploit item content (e.g., semantic) information in memory, or, may alter the way in which memories are able to be located in a given context. The following section describes an overview of a model framework from which two similar models are derived: the "Item Model" (IM), which assumes that testing strengthens the representation of item content, and the "Context Model" (CM), which assumes that retrieval causes additional context information to be represented in memories. The two models derive from the same framework, and thus are described together with the exception of their key differences that occur at the time of retrieval. Following an overview of the models, four experiments are reported, designed to test competing predictions from the two model variants.

Overview of the Model Framework

The testing effect model described below is based in REM theory (Shiffrin & Steyvers, 1997), and derived more specifically from a variant of REM from Malmberg and Shiffrin (2005), which combined aspects of REM and SAM (Raaijmakers & Shiffrin, 1981) to support a basic free recall paradigm. The model specifies that memories can exist in the form of lexical/semantic traces or as episodic traces. Throughout the course of an experiment, items that are presented to a participant to be learned are stored as episodic traces which include both content information (i.e., information about the item itself, such as semantic or perceptual features) and context information (i.e., information about the contextual state in which the item is presented). Content information for an episodic trace is stored as a lossy, error-prone copy of the lexical/semantic trace of an item being learned, whereas the context information in an episodic trace draws from the current state of context a participant is in, which presumably changes gradually over time.

During initial study, new episodic traces are established in memory drawing from the process described above. In a typical testing effect paradigm, items are again exposed during an intervening phase, either in the form of test trials or restudy trials. In the case of test trials, the model assumes that a search and recovery process is attempted to recall the item being cued (or in the case of free recall, any item previously encoded). Similarly, during restudy trials, an equivalent search and recovery process is engaged in (i.e., study phase retrieval occurs) with the item serving as a cue for its self. When retrieval is successful (for either test or restudy trials), an item is strengthened by augmenting the content and context information stored in the recovered trace. Importantly, the degree to which content (IM) or context (CM) information is strengthened is inversely proportional to the difficulty of recovery of the item, which in turn is dependent on the strength of the existing content stored in the episodic trace being recovered, along with the information present in the provided cue. The IM and CM differ in their assumptions about whether content (IM) or context (CM) information is strengthened by testing to a greater degree than restudy. Thus, restudy trials lead to little strengthening of episodic content or context information as such trials are performed with a very strong cue (i.e., the item its self), whereas test trials receive a more sizable increment to episodic content (IM) or context (CM) information as the cue used during a test trial- if present at all- is not a complete copy of the item being retrieved. In addition to the strengthening of episodic content or context information following successful retrieval, episodic images receive a baseline shot of additional content (CM) or context (IM) information (the latter following from Malmberg & Shiffrin, 2005), the size of which is not dependent on cue strength. In sum, successfully retrieved test and restudy items both receive similar sized baseline increments to content (CM) or context (IM) information (i.e., a single "shot"), but test items receive a larger increment to content (IM) or context (CM) information than restudied items, the degree of which is inversely proportional to the strength of the retrieval cue and the existing episodic image content.

Time is modeled as passing by a gradually drifting state of temporal context. As such, experimental delays (e.g., a retention interval) can be simulated by altering the current state of context (as represented by a vector of feature values) to a degree corresponding to the length of delay, or the magnitude of context change. Given that current context is utilized during retrieval, such context drift is able to alter recall performance as described in detail below.

Detailed Implementation of the Model

Initial study. Lexical/semantic images (i.e., the generic, stored representations of stimuli used in an experiment) are represented as vectors of features (holding *w* features, set to 20 for the present simulations) drawn from a geometric distribution with base rate $g = .45$. As such, lexical/semantic images contain feature values ranging from [1 - infinity), with higher values as increasingly rare, presumably representing more unique or distinct features of a given item. Similarly, current context is represented as a vector (size *w*) drawn from the same distribution. During initial study, episodic images are laid down item-by-item as error prone, incomplete copies of lexical/semantic images with associated context information. Following from Malmberg and Shiffrin (2005), the likelihood of storing a given feature increases with the number of storage attempts, *t*, such that a feature will be stored (i.e., a non-zero value stored in the episodic trace vector) with probability $1 - (1 - u)^t$. When storage of a feature occurs, there is a probability, *c*, that the value stored will be accurate (i.e., the same as drawn from the lexical/semantic image or the current context), otherwise a random value is drawn from a geometric distribution. As such, episodic images are both lossy (i.e., storage is not always successful), and error-prone (storage is not always correct).

The amount of time an item is studied dictates the amount of features stored in an episodic trace in a negatively accelerated fashion. That is, the number of storage attempts, *t*, at storing a piece of content information into an episodic trace can be described as $t_j = t_{j-1}(1 + e^{-aj})$ where *a* is a scaling parameter to adjust the rate of storage, t_1 is the number of storage attempts in the first one second of study, and *j* represents the study time in seconds. Context information is stored in a similar manner, with the exception that the amount of context that can be stored is capped, assuming that after two seconds, no additional information is stored (i.e., regardless of study time, the number of context storage attempts is capped at $t_{2,}$). The above described rules for content and context storage are entirely derived from Malmberg and Shiffrin (2005) and maintained for consistency; however they have little bearing on the issues under investigation in the present manuscript.

Free recall. The model implements two forms of retrieval: free recall and cued recall (the latter used for both cued recall test trials and for restudy trials). The two methods are largely similar, and thus free recall will be described first, with the additional assumptions then outlined for cued recall. During free recall, a cue-dependent search of all stored episodic traces in is carried out through a two-step sampling and recovery process. The search process described is based on a simplified version of that used in SAM (Raaijmakers & Shiffrin, 1981), using only the current state of context as a cue. The current context is utilized as a search probe, and matched to the context features of all episodic traces, with a likelihood ratio assigned to each image based on the relative match between the cue (current context) and the stored image context. Likelihood values are calculated such that matches in features between the cue and image increase likelihood, whereas mismatches decrease likelihood, with greater feature values weighted more heavily:

$$L_j = (1-c)^{n_{ij}} \prod \left(\frac{c+(1-c)g_s(1-g_s)^{i-1}}{g_s(1-g_s)^{i-1}}\right)^{n_{ijm}}, \qquad (1)$$

where g_s represents a base rate for the occurrence of feature values, i is a context feature value (1 to infinity), n_{ij} is the number of mismatching feature values in the episodic image, I_j, and n_{ijm} is the number of matches of feature i with value j. A specific image, I, is then selected probabilistically given the relative match between the cue (Q) and the image versus the match of the cue to all other images in memory:

$$P(I_i|Q) = \frac{L_i^y}{\sum L_k^y}, \qquad (2)$$

where y is a scaling parameter.

Once an image has been sampled, recovery is then attempted. Whereas the probability of sampling an item depends on the relative match between cue and an image, the probability of recovery depends on the absolute strength of the stored episodic image information. The probability of recovery is specified as:

$$P(R) = b(p_c^t * p_x^t) \qquad (3)$$

where p_c is the proportion of matching content features between the sampled episodic image and the target lexical/semantic image, p_x is the proportion of matching context features between the sampled episodic image and the current context, and b and t are scaling parameters. Thus, the likelihood of recovery of an item depends on both the absolute match between stored content features to the target, and stored context features to the current context. Note that this recovery method differs from the typical implementation of recovery in similar models (e.g., Malmberg & Shiffrin, 2005; Raaijmakers & Shiffrin, 1981) in that the degree of context match is taken into account (in the other models mentioned, recovery is based only on the strength of content information without regard to stored context information). The inclusion of context match during recovery is necessary to account for increasingly low levels of recall following variable (and often, long) retention intervals as are commonly used in the testing effect literature. Although a larger context mismatch will hurt the likelihood of sampling a specific item, this decrease rapidly asymptotes, given that even under circumstances of completely mismatching learning and test contexts, the likelihood of sampling a given item with a context cue equals 1/n, where n is the number of episodic images in memory, as sampling is dependent on the relative match between cue and images.[1] Including context match during recovery allows performance to drop to near floor following exceedingly long retention intervals given that recovery depends on the product of image vs. target content match and encoding vs. current context match. Conceptually, this characteristic of the model can be thought of as a participant needing to recall a specific item that was studied, along with the fact that it was encoded in the experimental context.

When free recall is initiated, participants engage in a repeating series of sampling and recovery attempts. First, an image is sampled. If the sampled item has not previously been output during the free recall procedure, recovery is attempted. If successful, the recovered item is output, incremented (as described below) and a new item is sampled (with replacement of the

[1] Mensink & Raaijmakers (1984) implemented a contextual drift mechanism into the SAM framework, including a means to uniformly reduce the likelihood of sampling *any* image by adding a positive, constant value to the denominator of the sampling equation. However, given that in testing effect paradigms there are always at least two sources of delay (between initial study and test(s), and initial test(s) and final test), typically of different durations, and that items may or may not have been exposed at various periods in an experiment (e.g., unsuccessfully tested items are not re-exposed whereas retrieved items are), such a mechanism of lowering sampling likelihood globally by a single constant was not used. Instead, in the present model, the effect of delay on recall manifests through the lower context match during both sampling and recovery following a delay.

previously recovered image) following the same procedure. If a previously output image is again sampled, or if recovery is unsuccessful, a recall failure occurs and a new item is sampled. Free recall terminates when K_{max} failures accumulate.

Cued recall. The above sampling and recovery procedure is also used during cued recall trials with a few exceptions. First, a cue is constructed as a lossy version of the lexical/semantic image of the target (i.e., a fragment of a given item serves as its own cue. That is, each feature of the target image is copied into a new vector with a probability defined by a parameter governing cue strength, *q*, otherwise no feature is stored. Sampling is conducted by calculating the relative match between the context cue and the context information stored in each episodic image, as done during free recall, in addition to the match between the item cue and episodic content information:

$$P(I_i|Q,C) = \frac{L_{ic}^{y} L_{ix}^{y}}{\sum L_{kc}^{y} L_{kx}^{y}}, \qquad (4)$$

where *L* refers to the likelihood as derived from equation 3 between the item cue, *C*, and episodic image content (subscript "c" likelihoods), or the current context cue, *Q*, and the episodic image context (subscript "x" likelihoods). Recovery is then attempted after an image, *I*, is sampled. The model assumes that all information available is used during recovery (i.e., the item content stored in the episodic image, in addition to the item content information provided by the cue its self). Thus the cue (which is a lossy copy of the target lexical/semantic image) is merged with the sampled episodic image content into a new "combined" vector, and recovery is then attempted following equation 3 using the combined vector to calculate content feature matches (p_c).

On a cued recall trial, the sampling and recovery process repeats in the same manner as during free recall, with the exception that a recovered item is only output if it is correct (this simplifying assumption was made for purposes of the present investigation, though a more complete model should account for cued recall output errors). The cued recall sampling and recovery process repeats until the target is successfully recovered and output, or C_{max} failures accumulate, where a failure occurs when an incorrect image is sampled or recovery fails.

Restudy. The model assumes that study phase retrieval occurs during re-presentations of any items during a simulation. The procedure mimics that of cued recall, with the key contrast being that the item re-presented serves as a cue for its own episodic image. That is, during cued recall cue construction as described above, the cue constructed during restudy is simply a copy of the lexical/semantic image studied with perfect fidelity (i.e., q = 1.0), and the sampling and recovery cycle proceed as in a cued recall trial. However, if study phase retrieval fails, a new episodic image is encoded as in the same manner as described during initial study, with the exception that context and content features are both encoded with a low fidelity (probability of storing a feature equals $u*t$), though note that the specifics of this component of the model are of little consequence to the present issues.

Incrementing. The key component of the model as it applies to the testing effect concerns, of course, the consequences of retrieval. However, two variations of the model are instantiated by altering the behavior of the model during incrementing: the IM (in which incrementing primarily impacts item content), and the CM (in which item context is emphasized). The IM is described first. Following successful recovery of a target (during free recall, cued recall, or study phase retrieval), the recovered episodic image is modified. First, a baseline "shot" of context is supplied to the recovered episodic image context vector (i.e., missing features are added) in the same manner as during initial study (though already stored features are not overwritten). In addition, missing features are added to the recovered episodic image content vector in the same manner as during initial study, but with probability:

$$P = (1-(1-u)^{t_j})(1-{p_c}^{m}) \qquad , \tag{5}$$

where m is a scaling parameter, and p_c, as in equation 3, refers to the proportion of features that match between the target lexical/semantic image vector and the combined episodic image content vector and item cue that was used during recovery. Thus, the stronger the combined episodic image plus item cue (or in the case of free recall, the episodic image content alone, as no item cue is provided) matches the target, the lower the likelihood of storing a new feature in the episodic image. Note that if the episodic image or combined episodic image and item cue perfectly match the lexical/semantic target image (i.e., p_c = 1.0) in which case the probability of storing a feature reduces to 0, the match value is reduced (to 0.99 in the present simulations) to allow some degree of incrementing to occur under such circumstances.

The CM behaves in a similar way to the IM model (described above), with the exception that the effects on item content and context are flipped. Item content is granted a baseline shot of information. Recovery of an item re-instates stored context information from the item into the current context (see, e.g., Howard & Kahana, 2002), with each element having probability *r* of updating. Importantly, following successful retrieval in the CM, item context is modified according to equation 5, such that the amount of item context strengthening is an inverse function of the degree of match between the retrieval cue(s) and item content. In other words, more difficult retrievals (i.e., weaker cues) yield larger increments to item context.

Context change. Delays in the model are implemented by randomly generating a new value for each feature in the current context vector with probability *D* (i.e., context randomly drifts), presumably with longer delays leading to more contextual drift, and thus, larger *D*. Multiple, variable delays occur in testing effect paradigms, but for present purposes, the contextual drift for simulating delay was only applied at the time of the retention interval, as described by the parameter D_{RI}. Contextual drift can also be useful for simulating other experimental tasks which, presumably, cause a shift in one's internal context state (e.g., distractor tasks, sources of interference, or list changes).

Overview of the Proposed Experiments

Four experiments were conducted to assess differing predictions of the IM and CM. Experiment 1 elaborated on work by Rowland et al. (2014), which examined the list-strength effect (Ratcliffe, Clark, & Shiffrin, 1990; see also Tulving & Hastie, 1972) as it relates to the testing effect. The IM and CM share differing predictions about the emergence of a list-strength effect in testing paradigms, given that the impacts of testing in the CM are primarily on the sampling process, whereas retrieval in the IM primarily impacts recovery. Past work in the area has suffered from non-perfect initial test performance, thereby making the results difficult to interpret given that retrieval success varied in test condition items. By boosting initial test performance to near ceiling, a more reliable assessment of the impacts of testing on item content versus context can be assessed.

Whereas Experiment 1 was designed to assess a key difference in the way in which testing impacts the sampling or recovery processes during recall, Experiments 2-4 examined the extent to which available contextual information is exploited at final test. In each Experiment, contextual information linked to the material learned in the initial study phase was either reinstated or not reinstated at the time of final test. Experiment 2 manipulated the reinstatement of an element of perception context (background color) from initial study. Experiment 3 manipulated semantic context cues, such that information was processed in relation to a weakly semantically related cue during learning, and cues were either reinstated or not reinstated at time of final test. Experiment 4 reinstated mental, temporal, context via an imagination task (following Sahakyan & Kelley, 2002), in which participations in a reinstatement condition were tasked with mentally recreating their thoughts in the moments preceding and leading into the initial study phase of the experiment. In the case of each experiment, the CM predicts that, during initial testing, contextual information from the initial study phase will be reactivated and stored within the item context image. In contrast, the IM model specifies that initial testing of an item will not lead to extra contextual information from the initial study phase to be stored into the item context representation. As such, the models differ in the extent to which context information from the learning episode is represented in the episodic images of tested items, and thus the extent to which that contextual information can be subsequently exploited to drive final recall. The IM model predicts that contextual information should not impact the magnitude of the testing effect, whereas the CM assumes a more effective exploitation of context cues, and thus a larger testing effect when such context is reinstated. Parameters used in the models to predict the general patterns of results examined in the experiments for the CM and IM are reported in Table 1.

Table 1. *Parameter values used in the simulations*

Parameter	Value	*Description*
w	20	(M&S) Size of all vectors (lexical/semantic, episodic content, episodic context, current context)
g	0.45	(M&S) Parameter used to sample a geometric distribution when generating feature values
g_s	0.40	(M&S) Parameter used in sampling likelihood ratio calculations
u	0.06	Probability of successfully storing a feature, per attempt.
t_1	6	(M&S) Number of storage attempts during first second of study.
a	1	(M&S) Scaling parameter for feature storage
c	0.95	Probability of accurately copying a stored feature
y	0.15	Scaling parameter for sampling likelihoods
t	2.2	Scaling parameter for recovery
b	26	Scaling parameter for recovery
m	10	Scaling parameter for incrementing
q	0.85	Probability of copying a target feature into a cue during cue construction
D_{RI}	0.3	Probability of drifting each current context feature at the retention interval
K_{max}	30	Number of errors at which free recall terminates
C_{max}	16	Number of errors at which cued recall terminates

Note. *(M&S) specifies the value used is that from Malmberg and Shiffrin (2005), many of which in turn were specified originally by Shiffrin and Steyvers (1997).*

EXPERIMENT 1

One method to examine the way in which information is represented in memory is to examine the extent to which changes to some items impact the ability to recall other items. In the context of the model, and the larger REM framework (see Shiffrin & Steyvers, 1997), there are two stages that are encountered during the task of recall: sampling and recovery. The sampling phase represents the process of locating a specific item in memory amongst other items, given a set of cues. More generally, sampling reflects the process of honing in and determining a specific item to retrieve from memory. Critically, the likelihood of a given item being sampled is a function of the *relative* match between available cues and the memory image. Manipulations that increase the match between a cue and target in memory will have the consequence of weakening the likelihood of sampling other, non-target items. In other words, increasing the match between a cue and a specific target also serves to decrease the match between that cue and other targets. As such, manipulations that impact the ability of the memory system to locate a given item will have an impact of making other, non-target items, less likely to be sampled. In recall tasks, context serves as a primary cue to guide the memory search. As such, items in memory that have a greater contextual match with the context cues used to guide retrieval will be more likely to be sampled, at the expense of other items that lack robust contextual information.

Once an item has been sampled (i.e., selected from amongst other items), the recovery phase refers to the task of unpacking the item for recall. Unlike sampling, recovery depends not on the relative match between current and stored context, but rather on the item content its self (i.e., recovery is a function of the *absolute* strength of the item content representation). As such, manipulations that impact item content, rather than context, will serve primarily to facilitate recovery, rather than sampling, in recall tasks.

The distinction between sampling and recovery has been particularly influential in memory modeling, as it allows for such models to account for list-strength effects, among other empirical observations (e.g., output interference, part-set cuing). The list-strength effect refers to the finding that certain types of manipulations that strengthen a subset of learned items also lead to a weakening of the remaining, non-strengthened subset of learned items (Ratcliffe et al., 1990). For example, in the case of the spacing effect when learning mixed lists consisting of both spaced and massed items, spacing serves to enhance recall of those items, but at the cost of weakening recall of massed items (Malmberg & Shiffrin, 2005). When spaced and massed items are learned separately (e.g., by using a between-participant manipulation), the magnitude of the effect tends to reduce. Recent work has observed a lack of a list strength effect in the testing effect (Rowland et al., 2014), with similar magnitude testing effects observed in both pure list (i.e., between-participant) and mixed list (i.e., within-participant, where lists consist of intermixed test and restudy items) designs. However, one problem with existing work concerns the lack of uniform treatment to test condition items. As elucidated by the bifurcation framework (Halamish & Bjork, 2011; Kornell et al., 2011), testing effect paradigms typically yield two classes of test condition items: those that are retrieved (and accrue whatever benefits that come from retrieval), and those that are not retrieved (and presumably are not strengthened or impacted as a result). Initial test data from Rowland et al. (2014) showed that participants successfully recalled approximately 75% of test items, yielding 25% of the test items unretrieved. When interpreting the results from the study, it is unclear whether the testing effect in fact does not yield a list-strength effect, or if testing yields either a positive or negative list strength effect that is then mitigated by the 25% of items that received differential treatment (i.e., unsuccessful initial retrieval) in the test condition. In fact, there is reason to suspect at least a possible masked list-strength effect from the Rowland et al. (2014) data, given that all restudy items were granted a spaced presentation (and thus received additional context; see Malmberg & Shiffrin, 2005), whereas only 75% of tested items had the opportunity for additional context to be stored following successful retrieval. One possibility is that test trials yielded a greater storage of context than restudy trials, but the difference in magnitude was counterbalanced by the smaller proportion of test items that received the extra context storage (75%), versus the 100% of restudy items. However, given the bifurcated test condition data, such a possibility is only speculative.

The IM and CM produce differing predictions about the list-strength effect that can only be reliably assessed under conditions in which nearly all test condition items receive the same treatment of successful retrieval. The IM model presumes that testing effects operate primarily through enhancing item content after retrieval, thereby making test condition items more likely to be recovered, but not impacting the sampling process. As such, the IM predicts a null list-strength effect, and instead a main effect of testing, given that in all conditions tested items enhance their absolute likelihood of retrieval (via impacting the recovery process). In contrast, the CM predicts a positive list-strength effect, such that testing serves to supplement item context information and thereby yield a greater likelihood of sampling tested items over restudied items. That is, the CM predicts that the impacts of testing will influence the sampling process, and thus make the tested items relatively more available to be located in memory, at the expense of less likely sampling of restudy items. Experiment 1 assessed the list-strength effect in a testing effect paradigm with a minimal bifurcation of test items, by adopting an initial test procedure that yielded high initial test performance.[2]

Method

Participants

One hundred participants were solicited using Amazon Mechanical Turk, drawing from a broad, online sample. Seven participants failed to complete the task (i.e., responses were not provided, or were off topic), and thus data were collected from 93 participants. There is recent work showing that online samples perform in a similar manner to traditional undergraduate participant pool samples, in a variety of tasks (e.g., Paolacci, Chandler, & Ipeirotis, 2010), including memory tasks similar in design to the proposed study (e.g., Rowland, Bates, & DeLosh, 2014). A power analysis indicated a required sample size of 90 in order to attain .8 power, assuming a testing effect size of *d* = .66 (based on the estimate from Rowland, 2014). Approximately 65% of participants identified as female, and ages ranged from 22 to 62.

Design and Materials

The experiment utilized a design following the method detailed by Erlebacher (1977), which is useful for examining list-strength effects. Participants were divided into three conditions: pure test; pure restudy; and mixed. Using three groups, Erlebacher's design allows for the examination of two factors: the manipulation of test versus restudy, and the nature of the manipulation (i.e., between vs. within participants). Stimuli to be learned consisted of sixteen unrelated nouns drawn from Wilson's (1988) MRC Psycholinguistic Database, constrained to 1-

[2] An alternative method to reduce bifurcation of test condition items is to provide feedback after each test trial (see Kornell et al., 2011; Rowland & DeLosh, 2014a), thus re-exposing testing condition items that are not successfully retrieved. This method was not employed, however, under the assumption that unsuccessful test trials followed by feedback would seem unlikely to reinstate prior context, if context reinstatement is indeed a mechanism contributing to the testing effect.

2 syllables, 6-8 letters, word frequency between 10 and 25 occurrences per million, and concreteness and imagability values between 300 and 500.

Procedure

Participants began by being presented with instructions of the task, including an indication that a list of words would be presented and should be remembered for a later memory assessment. First, during the initial study phase, 16 items were presented, sequentially, for 4 s each. Following the initial study phase, participants completed a 15 s mental math task, requiring them to complete a series of simple arithmetic operations and input their answer. After the short distractor task, participants were re-exposed to the stimuli, in a format depending on their condition. Those participants in the pure restudy condition saw all 16 items, sequentially in a new random order, for 7 s each, and were instructed to copy each item as shown by typing it into a provided text box. Participants in the pure test condition received fragments of the previously learned stimuli in a new random order, for 7 s each, and were instructed to recall and input the previously studied item that completed each fragment. Fragments were constructed in such a way that only 2 letters were removed from each item in order to promote a high level of initial test performance (e.g., "CA_CHE_" for the item "CATCHER"). Note that Rowland and DeLosh (2014a) used a similar paradigm and cuing procedure and were successful in yielding near-ceiling initial test performance. Participants in the mixed condition received 8 trials of test items and 8 trials of restudy items, randomly intermixed, and in the same format as the corresponding trials in the pure test and pure restudy conditions, respectively.

Following the intervening test or restudy phase, participants completed a 4 min distractor task (reporting as many world countries as they could generate). A free recall test lasting 90 s was then administered, in which participants were instructed to recall as many items as possible from the entire experiment. In total, the experiment lasted approximately 12 min.

Results and Discussion

Data are reported in Figure 1. No differences were found on initial test performance between the mixed (85%) and pure (83%) test groups, *t*(30) = .71, p = 48. Final test data were submitted to Erlebachers' (1977) ANOVA model, in which the effects of testing versus restudy, and of mixed versus pure lists, can be examined. A main effect of intervening task emerged, with tested items recalled at a higher frequency than restudied items, $F(1,89) = 4.75$, $p < .05$, $\eta_p^2 = .14$. The main effect of list type was not significant, $F(1, 64) = 0.02$, $p > .10$, $\eta_p^2 = 0$. A trend towards an interaction between intervening task and list type was observed, $F(1, 89) = 3.77$, p = .06, $\eta_p^2 = .11$ with the pattern of results indicating a list strength effect, such that a testing effect emerged in the mixed list condition, $t(30) = 3.93$, $p < .01$, $d = .69$, but not the pure list condition, $t < 1$, $d = .04$.

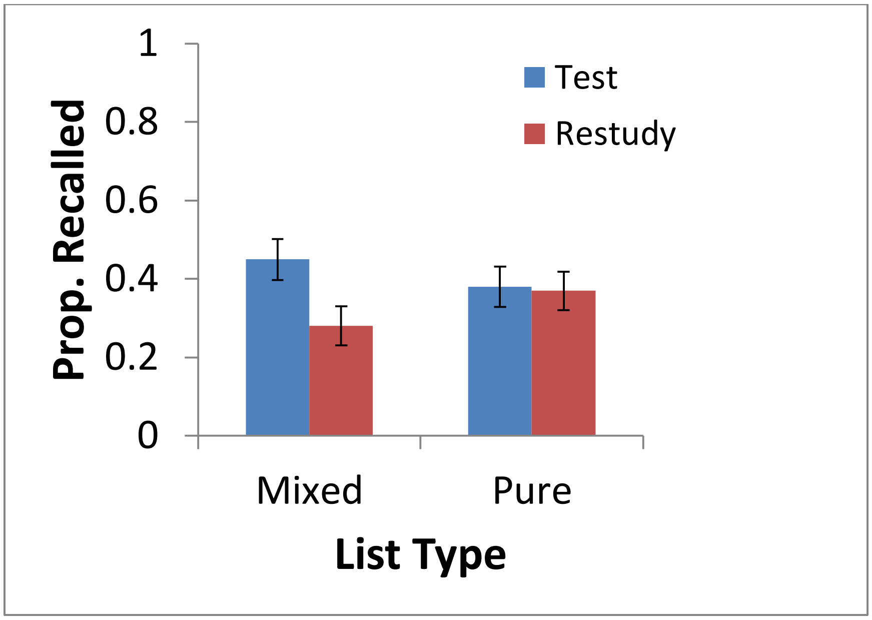

Figure 1. Experiment 1 Final Test Performance. Error bars represent standard error.

The results of Experiment 1 are compatible with the prediction of the CM model, which assumes that testing strengthens the representation of context stored in the memory trace. If the strong trend towards an interaction (*p* = .06) is interpreted as significant, a list-strength effect emerged, with a testing effect obtaining in the mixed- but not pure-list condition. This result is at odds with Rowland et al. (2014), who did not observe an interaction between list type and the testing effect. However, in addition to different levels of initial test performance, one key difference between the reported experiment and those of Rowland et al. (2014) involves the use of a single list in the present experiment versus a series of lists (i.e., interpolated study-test learning phases across multiple lists) in the past research. Although it is not clear what the precise influence of a single vs. multiple list design on the testing effect is, it is possible that any effects of testing that influence the storage, or recovery of context (e.g., see Howard & Kahana, 2002), may interact with the presentation of segregated lists. For example, multiple list designs may yield less interference in a pure test versus pure study condition if testing serves to recover and contextually segregate the separate lists. Similarly, compared with the previous work, participants in the present study were exposed to a smaller number of stimuli (16, versus 36 total items in Rowland et al.), and thus if testing serves to restrict the search set used at final test, it may serve to be more beneficial to memory under circumstances where a larger set of stimuli are learned, regardless of the type of list employed. Even so, such possibilities are speculative and may warrant further investigation. For present purposes, the key implication of the present results is that they suggest that testing may serve to strengthen the representation of context in a memory trace, as inferred by the observation of a list-strength effect (see Malmberg & Shiffrin, 2005).

EXPERIMENT 2

Experiment 1 investigated whether testing primarily influences the memory systems' ability to search for an item in the sampling process (as influenced primarily by context information), or to unpack an item after it has been located in the recovery process (as influenced primarily by content information). In contrast, Experiment 2 (in addition to Experiments 3 and 4) is designed to examine whether context information can be effectively utilized during final recall to help locate information previously learned via testing. This issue was examined by presenting information to be learned in the presence of unique contextual cues during initial study. At a final free recall test, context cues associated with the initial study phase are either re-presented or absent. The CM assumes that, during initial retrieval, contextual elements from initial study are re-instantiated and subsequently strengthened within the successfully retrieved test item episodic images. As such, the explicit re-instatement of those contextual elements at the time of final test is predicted to facilitate the sampling and subsequent recall of test items. As such, the CM predicts that context cues at final test should increase the magnitude of the testing effect. In contrast, the IM assumes that testing enhances item content, not context. Thus, the explicit reinstatement of contextual cues at final test should similarly (rather than differentially) influence the sampling of test and restudy items, and as such the testing effect is predicted to be of similar magnitude with or without final test context cues provided. In Experiment 2, the method of manipulating context (specifically, perceptual context) cues was adapted from Isarida and Isarida (2007), in which material is studied in the context of either of two different background colors. At final test, one of those background colors was reinstated, thus providing a manipulation of the contextual cues available at test (i.e., the same color as during initial study, or a different color from initial study).

Method

Participants

Forty-five participants were solicited via Amazon Mechanical Turk. Five participants were dropped for failing to provide valid responses, yielding a useable sample size of 40. A power analysis indicated a required sample size of 34, given effect size $f = .25$, and .8 power. Effect sized estimates were derived from assuming a medium effect, similar to that observed by Isarida and Isarida (2008), from which the context manipulation was adapted. Approximately sixty percent of participants identified as female, and ages ranged from 19 – 53.

Design and Materials

The experiment employed a 2x2 within-participant design, manipulating intervening task (test versus restudy) and final test cue reinstatement (same versus different). A stimulus set of 16 unrelated nouns was generated using Wilson’s (1988) database, following the same constraints as used in Experiment 1, with the exception of increasing the word frequency range to 25 – 75 occurrences per million.

Procedure

Participants began by entering an initial study phase in which the 16 words were presented, sequentially, for 4 s each. Each word was presented against either a blue or red background color. The order of item presentation (and thus the background colors) was random, with the constraint that each background color was presented an equal number of times, and no more than two consecutive trials passed in which the same background color was presented.[3] Following initial study, a 15 s mental math distractor was administered, after which participants completed the intervening task phase of the experiment. The same method used in the mixed list condition of Experiment 1 was employed (i.e., 7 s per item, with test and restudy trials randomly intermixed, user responses solicited and typed in, and test items cued with fragments missing 2 letters), and with the constraint that equal numbers of both the test and restudy items were associated with each the two background colors from the initial study phase. Thus, 8 items were tested and 8 restudied. The trials during the intervening task were presented against a neutral (white) background color.

Following a 3 min distractor task, a final free recall test was administered for 90 s. During the final test, participants received a context reinstatement of one of the background colors exposed during initial study (i.e., blue or red). In sum, an equal number of items belonged to each of four crossed conditions: tested same context (e.g., studied with a blue background, tested, and with a final test blue background); tested different context; restudied same context; and restudied different context.

Results and Discussion

Data are presented in Figure 2. Initial test performance between the same (84%) and different (79%) reinstatement conditions did not significantly differ, $t(39) = 1.48$, $p = .15$. Final test data were submitted to a 2x2 repeated measures ANOVA, with intervening task (test versus restudy) and cue reinstatement (same versus different) as factors. The main effect of intervening task was significant, with tested items recalled at a higher frequency than restudied items, $F(1, 39) = 17.47$, $p < .01$, $\eta_p^2 = .31$. However, neither the main effect of cue reinstatement,

[3] Isarida and Isarida (2007) found that randomly intermixing colors yields more robust context effects than blocking items by background color.

F(1, 39) = .027, *p* = .87, η_p^2 = 0, nor the interaction, *F*(1, 39) = 1.01, *p* = .32, η_p^2 = .03, reached significance.

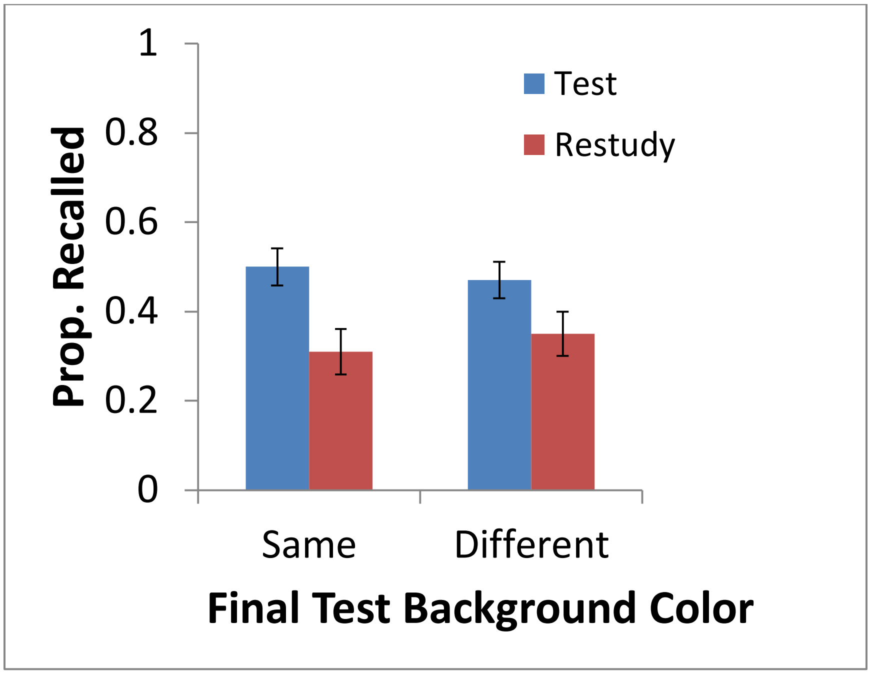

Figure 2. Experiment 2 Final Test Performance. Error bars represent standard error.

Although the results of Experiment 1 suggested that testing likely impacts context memory, Experiment 2 did not find statistically significant support for the CM model, suggesting that context reinstatement, at least at the perceptual, item-specific level, does not largely influence the testing effect. Note, however, that the numerical trends were in support of the predictions of the CM. That is, there was a .19 mean difference between tested and restudied items in the same context reinstatement condition (*d* = .54), whereas the difference was only .12 in the different context condition (*d* = .4). Even so, this pattern was not statistically significant. In contrast, the results are consistent with the IM model, which assumes that testing serves to strengthen item content, and thus should enhance memory similarly with regard to the presence or absence of reinstated contextual elements.

EXPERIMENT 3

Experiment 2 examined the effect of a specific type of context reinstatement: perceptual. An additional form of context concerns the conceptual, or semantic, context in which a given piece of information is processed. A method that can be used to examine this type of context involves manipulating the types of cues that are presented. For example, cuing the target "chair" with "table," leads to processing the target in a different semantic context when compared with the cue "committee." Thus, a semantic cue can serve as a contextual element peripheral to the target. Experiment 3 thus adapted a method, derived from Thomson and Tulving (1970), in which participants were exposed to target items paired with weakly associated contextual cues. Following a test or restudy opportunity, a final cued recall test was administered with items cued either in the presence or absence of the weakly associated contextual cue.

Such a manipulation fits within the domain of the encoding specificity principle (see also, Morris, Bransford, & Franks, 1977). In brief, memory is often the most accurate when the conditions at retrieval match the conditions at encoding (though cf., Nairne, 2002). As such, reinstating a semantic cue to guide retrieval of a given target at the time of final test can be considered to be reinstating a form of context that relates to the conceptual nature of the episodic event, rather than a perceptual element. A primary purpose of Experiment 2 is thus to examine the possibility that the relationship between testing and context may be selective as to the specific type of context under consideration.

Similar to Experiment 2, the CM predicts that context reinstatement should increase the magnitude of the testing effect. As such, an interaction between the effects of testing and context is predicted to emerge, with a larger magnitude testing advantage in the context reinstatement condition. In contrast, the IM assumes that context reinstatement should have no preferential effect on tested or restudied information, and thus the model predicts a main effect of testing, but without an interaction. Thus, the key difference in predictions between the two models concerns the presence (CM) or absence (IM) of an interaction between testing and context reinstatement.

Method

Participants

Fifty-five participants were solicited from Amazon Mechanical Turk, with 8 participants failing to complete the task, thus yielding data from 47 participants. A power analysis indicated a required sample size of 34, based on the parameters employed for Experiment 2. Given that the models do not differentiate between specific types of context, the same effect size to that used in the Experiment 2 power analysis was estimated. Reported females made up approximately 70% of the sample, and ages ranged from 20 – 67.

Design and Materials

Two factors were manipulated within-participants: intervening task (test versus restudy), and cue reinstatement (reinstated versus not reinstated). Sixteen nouns, serving as targets, were generated from the MRC Psycholinguistic database (Wilson, 1988), constrained to 5-7 letters, 1-2 syllables, a frequency range of 100 – 225 occurrences per million, and concreteness and imagability values of 500-700. Word frequency, concreteness, and imagability were increased from the parameters used in Experiments 1 and 2 in order to ensure that participants would be able to more easily interpret the target in the context of an associated cue. Each of the cues were weakly associated to the targets (forward associative strength = .01), and derived from the USF Free Association norms (Nelson, McEvoy, & Schreiber, 1998).

Procedure

Participants were instructed that they would be exposed to word pairs, with a target presented in upper case, and a cue in lower case. They were instructed to learn the target words by thinking about how they relate to the cue words. During initial study, each word pair was presented for 5 s, sequentially, in the form “cue – TARGET.” After the 16 initial study trials, a 15 s distractor ensued, followed by the intervening task. A random half of the items were tested, with the other half restudied, in a randomly intermixed order, in the same format as Experiments 1 and 2 (i.e., the target was presented either in full or missing 2 letters). A 3 min distractor task followed the initial study phase, and subsequently, a final cued recall test was administered. During the final test, participants were presented, sequentially, with the first two letters of each of the 16 targets serving as cues (e.g., “TA_____”, for the target TABLE). In addition, a random half of the target cues were accompanied by the weak associate context cue (e.g., “wood – TA____”). Participants had 10 s to attempt retrieval before moving to the next item. The experiment lasted approximately 12 minutes.

Results and Discussion

Data are presented in Figure 3. Performance on the initial test did not differ between reinstated (78%) versus non-reinstated (81%) items, $t(46) = .60$, $p = .55$. Final test performance was examined using a 2x2 repeated measures ANOVA. Main effects of intervening task, $F(1,46) = 8.93$, $p < .01$, $\eta_p^2 = .16$, and cue reinstatement, $F(1,46) = 10.42$, $p < .01$, $\eta_p^2 = .19$ were significant, with tested items recalled at a higher frequency than restudied items, and reinstated cues leading to better performance than non-reinstated cues. No interaction was observed between the two factors, $F(1,46) = .07$, $p = .80$, $\eta_p^2 = 0$.

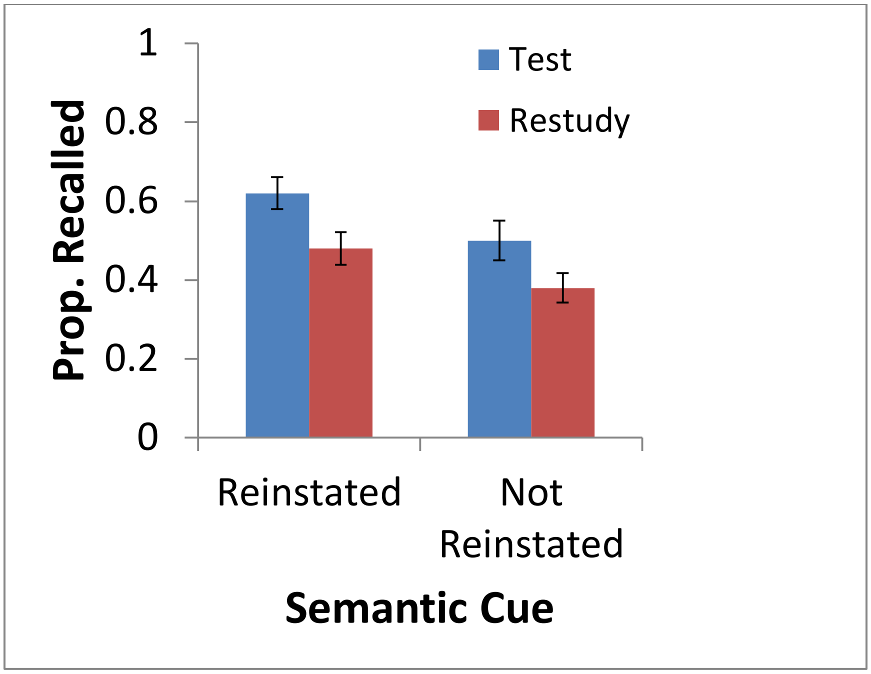

Figure 3. Experiment 3 Final Test Performance. Error bars represent standard error.

Similar to Experiment 2, testing was beneficial for memory overall, though there was no influence of cue reinstatement on the magnitude of the testing effect. As such, the testing effect does not appear to be reliably influenced by reinstating either perceptual or conceptual context cues at the time of final test.

EXPERIMENT 4

Whereas Experiments 2 and 3 examined the effect of perceptual and semantic context reinstatement on the testing effect, Experiment 4 was designed to assess mental, temporal context reinstatement. To accomplish this goal, a mental imagination and reinstatement task was adapted from Sahakyan and Kelley (2002). Using this method, a context change can be induced by invoking a mental state in participants that differs drastically from their current experimental task. For instance, Sahakyan and Kelley had participants mentally walk through and describe their parents' home, or describe what they would do given the ability to be invisible. Later, context can be reinstated by having participants explicitly recount the thoughts and feelings they had near the beginning of the experiment. As such, using an imagination task allows a different type of context to be reinstated when compared with the perceptual and conceptual elements of Experiments 2 and 3; namely, mental, or temporal context.

An additional deviation of Experiment 4 from Experiments 2 and 3 concerns the breadth of association between the context cue and target items. In Experiments 2 and 3, the reinstated context cue varied rapidly, trial to trial (i.e., background colors were randomly intermixed, and semantic cues were unique to each target). However, the mental context reinstated in Experiment 4 was stable during learning, and common to all the stimuli learned by the participant. That is, all of the targets were associated with the same list context. The importance of this distinction is elaborated upon in the General Discussion.

As in Experiments 2 and 3, the key finding of interest concerning the models is the presence or absence of an interaction between testing and context reinstatement. The CM predicts an interaction, such that the testing effect should be larger under conditions of context reinstatement. In contrast, the IM predicts no influence of context reinstatement on the magnitude of the testing effect, and as such, predicts that only a main effect of testing should emerge.

Method

Participants

Eighty-two participants were solicited via Amazon Mechanical Turk. Ten participants failed to follow task instructions, yielding data from 72 participants. A power analysis indicated a target sample size of 68 participants (power = .8; *f* = .25*)*, again, based on the assumption of a similar effect size given that the models do not differentiate between types of context. Approximately 60% of participants identified as female, and ages ranged from 19 – 64.

Design and Materials

Experiment 4 employed a 2x2 mixed design, with intervening task (testing vs. restudy) manipulated within-participants, and context reinstatement (reinstated vs. not reinstated) manipulated between participants. 36 participants were assigned to the reinstatement condition, and 35 to the non-reinstatement condition. Sixteen stimuli were generated using the same parameters as Experiment 2.

Procedure

The procedure was similar to that of Experiment 2 in most aspects. During initial study, each of the 16 items was sequentially presented for 4 s each. After a 15 s math distractor task, a randomly intermixed 8 items were given a restudy opportunity, and the remaining 8 items were initially tested, with 7 s per trial. After the intervening task, all participants received an imagination task designed to promote a mental context change.

The task was derived from Sahakyan and Kelley (2002, Exp. 2): participants were given 90 s to mentally "walk through" and describe the layout of their parents' home. Descriptions were solicited via a text box. Next, those participants in the no-reinstatement group were given a 90 s distractor (generating and reporting animal species), whereas participants in the reinstatement group were given a 90 s experimental context reinstatement imagination task. The task, also derived from Sahakyan and Kelley (2002, Exp. 2), requested participants to recall and imagine the thoughts, feelings, and emotions that they experienced immediately preceding the experiment, and furthermore, to consider any thoughts, feelings, or emotions that arose as the experiment began. Descriptions were solicited via a text box.

All participants were next given a 90 s free recall period, in which they were asked to recall as many of the originally studied items as possible. The experiment lasted approximately 12 minutes.

Results and Discussion

Data are presented in Figure 4. Participants did not differ in initial test performance between the reinstatement (83%) and no-reinstatement (79%) groups, $t(69) = .74$, $p = .46$. Final recall data were submitted to a 2x2 mixed factor ANOVA. A main effect of intervening task was detected, with tested items recalled at a higher frequency than restudied items, $F(1, 69) = 21.68$, $p < .01$, $\eta_p^2 = .24$. The main effect of context reinstatement was not significant, $F(1, 69) = .05$, $p = .83$, $\eta_p^2 = 0$. However, an interaction between the two factors was significant, $F(1, 69) = 4.30$, $p = .04$, $\eta_p^2 = .06$ with a larger testing advantage observed in the non-reinstatement compared with the reinstatement condition. Post-hoc means comparisons showed that a significant testing effect emerged in the non-reinstatement condition, $t(34) = 4.66$, $p < .01$, $d = .81$, whereas in the reinstatement condition, the testing effect did not reach significance, *t*(35) = 1.87, *p* = .07, *d* = .32.

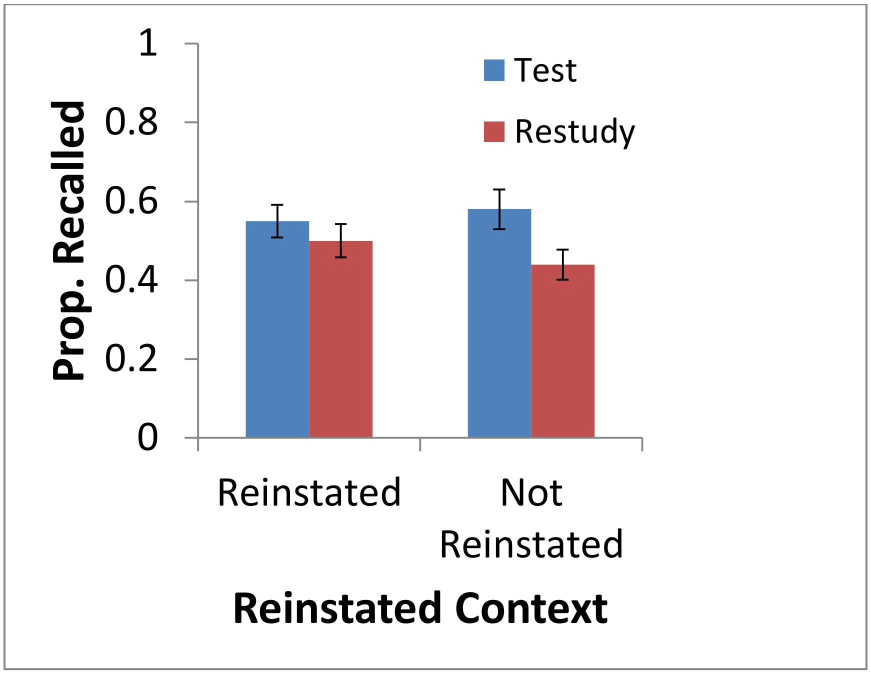

Figure 4. Experiment 4 Final Test Performance. Error bars represent standard error.

Experiment 4 differed from Experiments 2 and 3 by demonstrating an influence of the reinstatement of a contextual element on the testing effect. However, in contrast to the prediction of the CM, the testing effect was in fact larger in the non-reinstatement condition, indicating that testing may be more beneficial when in the absence, rather than presence, of explicitly reinstated context.

GENERAL DISCUSSION

The testing effect was examined across a series of four experiments designed to test competing predictions of models that rest on differing assumptions about how retrieval alters memory. The IM assumes that the act of testing serves to strengthen the item content stored in a memory trace, thus leading tested material to be more easily recovered after it is located in the memory search process. In contrast, the CM assumes that retrieval enhances the storage of contextual elements in memory. In particular, retrieval is assumed to cause a reinstatement of prior context, which is then updated into the memory trace along with the current state of context during the retrieval trial. As such, the CM specifies that testing serves to alter the ability of the memory system to sample a specific item amongst the other items in memory. By adding additional diagnostic information to the memory trace, the CM specifies that testing makes retrieval more likely in large part by increasing the utility of context cues when sampling possible candidates for retrieval (see Raaijmakers & Shiffrin, 1981). In short, the CM model assumes that testing helps one locate an item in memory, whereas the IM model assumes that testing helps strengthen an items' content directly.

Results from the experiments lend mixed, equivocal support to the models. Experiment 1 revealed a list-strength effect, with a testing effect emerging in mixed- but not pure- lists, consistent with a prediction of the CM. More specifically, when tested and restudied items are intermixed in the memory system, initial testing appears to lead to a greater selective sampling of tested items during a final free recall test. Because sampling operates according to the relative match between cues available at test (i.e., the current state of context, in the case of free recall) and information stored in memory, a list strength effect indicates a greater degree of context match between the final free recall context cue and the contextual information stored in tested memory traces (see also Malmberg & Shiffrin, 2005). Thus, although Experiment 1 indicates that testing likely interacts with context memory in some way, it does not elucidate the specific nature of the interaction. Experiments 2-4 serve as a way to examine the relationship between testing and context in more detail.

Experiments 2 and 3 did not observe statistically significant effects of context reinstatement on the testing effect, consistent with the IM. That is, reinstating background color (Exp. 2) did not alter the testing effect, nor did reinstating semantic cues (Exp. 3). However, of particular note, Experiment 4 revealed an influence of context reinstatement on the testing effect, though in contrast to the predictions of the CM, explicit context reinstatement in fact mitigated, rather than enhanced, the magnitude of the testing effect. As such, differences in the nature of the context that was reinstated in Experiment 4, compared with Experiments 2 and 3, are of particular interest.

Context, as described in the Introduction, can refer to a wide variety of elements. In the present study, the contextual elements examined include perceptual, semantic/conceptual, and mental/temporal features, and thus are themselves qualitatively different from each other. One salient distinction between types of contextual features, as described by Glenberg (1979), concerns whether a given element of context varies rapidly, trial-by-trial (i.e., “local” context), or stays relatively stable across trials (i.e., “global” context). Experiments 2 and 3 both manipulated the reinstatement of local contexts: in Experiment 2, background color randomly changed, trial by trial, during learning, and was item specific. In Experiment 3, semantic cues were unique to each target. In contrast, the reinstated context in Experiment 4 was global and common to all items: the internal, mental and temporal context associated with the entire learning phase of the experiment. This distinction might serve as a useful way by which to consider the discrepant results across Experiments 2-4. Specifically, two issues are discussed: the observation that testing seems to interact with manipulations of global contexts, but not local contexts, and secondly, the nature of that interaction, whereby the testing effect is larger in the absence, rather than presence of explicitly reinstated contextual cues.

Little existing work has examined the relationship between testing and local contextual elements, and even then, there is not any clear consistency. Rowland (2011) observed a statistically significant, albeit small, effect of testing on enhancing later memory for perceptual elements of local context (i.e., font color of item presentation). However, Brewer et al. (2010) failed to find an influence of testing on memory for varied acoustic information present at learning (the gender of a speaker reading stimuli). In contrast to local context, most work that is directly relevant to the relationship between testing and context has concerned contexts that are both shared by many items and are relatively slow to change. As described in the Introduction, testing has been found to enhance list-discrimination (e.g., Chan & McDermott, 2007), such that participants are better able to discern which- of multiple- lists a given item belongs to. As such, one possibility is that the contextual information updated in memory following retrieval largely consists of slowly changing and global contextual information. Indeed, this type of gradually drifting context (e.g., temporal context) is often emphasized in models of memory, including the Temporal Ratio Model (Brown, Neath, & Chater, 2007), the Temporal Context Model (Howard & Kahana, 2002), and similar models such as the Context Maintenance and Retrieval Model (Polyn, Norman, & Kahana, 2009). Relatively less attention has been drawn towards elements of rapidly changing local contexts.

In terms of developing the present models, future work may seek to examine the effects of different types of context, and their representations, within the models, with the distinction between local and global contexts acting as an initial step. Indeed, the discrepant results of Experiments 2 and 3, compared with Experiment 4, indicate that the effect of explicitly reinstating context depends on the nature of the context being reinstated. The model framework used in the present study, however, treats context as an undifferentiated construct, and in the present implementation, contextual elements may drift gradually (e.g., time), or rapidly (e.g., trial-to-trial background colors). One possibility is to view context as solely temporal, or as the gradually drifting mental state. In this view, elements of "context" that change rapidly and are linked to items themselves (e.g., background color, semantic cues), may better be represented as item content, rather than context information. Benjamin (2010) argues that elements peripheral to the focus of attention in a memory task are not represented as "context," but rather as content (the same as target information), though with a lower resolution. One possible modification of the IM and CM could be to represent item content in such a way that all elements that change rapidly, trial by trial (i.e., local contexts), are represented, and with "context" reserved only for the gradual passage of time or mental state. This construction of the models may provide a means by which to interpret the observed results of the experiments. That is, no effects of context reinstatement on the testing effect were observed in Experiments 2 and 3, perhaps because *context* (i.e., global, temporal, mental context) was in fact not manipulated. However, Experiment 4 did observe an interaction, perhaps because it provided the only true manipulation of global context in the study. Such possibilities may be useful to explore further.

Given the emphasis of global contexts in memory modeling, the relationship between global context reinstatement and the testing effect may prove easier to reconcile within a memory modeling framework. The CM model assumed that context reinstatement would be exploited by tested items, and thus increase the magnitude of the testing effect. However, global context reinstatement may be redundant, rather than particularly exploitable, with the inherent effects of testing. This view is compatible with work from Szpunar et al. (2008), who found that testing, more so than restudy, helps one hone in on a specific subset of learned information (e.g., a specific list) in the absence of strong, explicitly provided context cues. That is, the act of testing, on its own, serves a function similar to that of explicit context reinstatement. Similarly, Rowland and DeLosh (2014b) found that testing a subset of items on a list can, under some circumstances, facilitate memory for other items from the same list context. In particular, Rowland and DeLosh claimed that testing may help reinstate prior list contexts, facilitating access to all of the information learned within such contexts. As such, although there have not been unequivocal, a priori predictions inferable from past work about the relationship between testing and context reinstatement, existing research is at least consistent with the possibility that testing inherently serves a function of reinstating contexts, and thus may be redundant with- rather than able to exploit- any explicit, experimenter-provided presented contextual cues.

An alternative but not mutually-exclusive possibility to explain the discrepant results across experiments is to assume that testing emphasizes the storage of only specific subtypes of context in memory. That is, testing may selectively enhance and utilize mental/temporal contextual information, while influencing other types of context (e.g., perceptual) to a smaller degree, if at all. Indeed, Brewer et al. (2010) found that, although testing enhanced list discrimination (see also, Chan & McDermott, 2007; Verkoeijen et al., 2011), it did not enhance memory for different aspects of context memory. In particular, in one experiment, Brewer et al. (2010) had participants learn stimuli by listening to items presented in either a male or female voice. Testing was not found to strengthen memory for the voice in which an item was learned in.

Indeed, the models introduced in the present study do not explicitly identify or discriminate between different types of features stored as item context. This ambiguity of the models is common to most work in the larger REM framework (Shiffrin & Steyvers, 1997), along with most other memory models that characterize memories as vectors of feature values (e.g., Hintzman, 1988; Howard & Kahana, 2002). In fact, little work has been done to even explicitly describe the nature of “features” stored in memory traces. Underwood (1969) outlined a taxonomy of possible types of information that are likely represented by features in memory traces (e.g., attributes concerning temporal, spatial, orthographic, and a variety of other characteristics about encoded events). However, there lacks any application of such fine grained distinctions of feature types into modern memory models, nor is there any clear indication of what aspect of a given feature classification is tapped by a feature value in a vector (e.g., if a feature value is assumed to represent an orthographic feature of the stimulus, what is the particular orthographic feature being represented?). That is, it is not clear what specifically *about* semantic content, or temporal context, is represented by a given feature value. As such, there exists an explanatory gap between our assumptions about elements of events in memory and their representations in models. Future work may benefit by adding more specificity to such representations.

Conclusions

Support for the models was mixed. The IM derived support from Experiments 2 and 3, but not 1 or 4. Experiment 1 was consistent with the CM. The outcome of Experiment 4 was not predicted by either the CM or the IM, though the pattern of results does suggest that testing interacts, in some form, with context. Taken together, the reported experiments provide mixed evidence as to the types of strengthening in memory that result from retrieval. The list-strength effect observed in Experiment 1 is consistent with the assumption that testing alters the search process in memory, likely by strengthening the representation of context in some form. Experiments 2-4, considered as a whole, suggest that testing may operate in part by helping implicitly reinstate global context features or mental/temporal features, to a greater degree than other, local contextual information. Thus, a tentative conclusion from the present study is that testing likely strengths item content in memory to some extent, in addition to facilitating the search process by implicitly reinstating certain context cues during recall. That is, testing likely recovers context on its own.

More generally, a contribution of the present work is to help elucidate the possible mechanisms through which testing strengthens memory. Despite an inconsistent pattern of results across experiments, it does seem clear that testing, in some fashion, operates at least in part by interacting with context. As such, future characterizations of the testing effect may benefit by considering the joint impact of item content strengthening, in conjunction with context strengthening and recovery. However, the specific nature of that interaction, and the types of context involved, remain areas of interest for further theoretical examinations of the testing effect.

REFERENCES

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

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396–401). New York: Wiley.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Meltion & E. Martin (Eds). *Coding processes in human memory*. Washing DC: Winston.

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

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

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576.

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

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

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

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

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin and Review, 19*, 443-448.

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

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

Congleton, A., & Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolished collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General, 140,* 535-551.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin, 84,* 212-219.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition 7,* 95 – 112.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 801-812.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528-551.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269-299.

Isarida, T., & Isarida, T. K. (2007). Environmental context effects of background color in free recall. *Memory & Cognition, 35,* 1620-1629.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649-667.

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

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85-97.

Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory, 19,* 346-359.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 322-336.

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

Mensink, G. & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434-455.

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

Morris, P. E., Fritz, C. O., Jackson, L., Nichol, E., & Roberts, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning and imagery. *Applied Cognitive Psychology, 19,* 779-798.

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation

Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory*, *20*, 138-154.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgement and Decision Making, 5*, 411-419.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116,* 129-156.

Potts, R., & Shanks, D. R. (2012). Can testing immunize memories against interference? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1780-1785.

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

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

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93-134.

Ratcliffe, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 163-178.

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review, 24,* 419-435.

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

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

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

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

Rowland, C. A. (2011). *Testing effects in context memory*. (Masters Thesis, Colorado State University).

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

Rowland, C. A., Bates, L., & DeLosh, E. L. (2014). On the reliability of retrieval-induced forgetting. *Frontiers in Psychology – Cognition, 5,* 1343.

Rowland, C. A., & DeLosh, E. L. (2014a). Mnemonic benefits of retrieval practice at short retention intervals. *Memory, 23,* 403-419.

Rowland, C. A., & DeLosh, E. L. (2014b). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin and Review, 21*, 1516-1523.

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition*, *42*, 912-921.

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

Sensenig, A. E., Littrell-Baez, M. K., & DeLosh, E. L. (2011) Testing effects for common versus proper names. *Memory*, *19*, 664-673.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin & Review, 115*, 893-912.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 260-271.

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

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86*, 255-262.

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

Tulving, E., & Bower, G. H. (1974). The logic of memory representation. In G.H. Bower (Ed.), The Psychological of Learning and Motivation (Vol. 8, pp. 265-301). New York: Academic Press.

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92,* 297-304.

Underwood, B. J. (1969). Attributes of Memory. *Psychological Review, 76,* 559-573.

Verkoeijen, P. P. J. L., Bouwmeester, S., & Camp, G. (2012). A short-term testing effect in cross-language recognition. *Psychological Science, 6,* 567-571.

Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology, 58*, 490-498.

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic Bulletin & Review, 18*, 518-523.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioral Research Methods, Instruments, and Computers*, *20*, 6-11.

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