Journal of Memory and Language 103 (2018) 19–27

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Testing enhances memory for context

Melisa Akan[a,*], Sarah E. Stanley[a], Aaron S. Benjamin[a,b]

[a] University of Illinois at Urbana-Champaign, United States
[b] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Keywords:
Testing effect
Context memory
Episodic retrieval

ABSTRACT

The beneficial effect of retrieval practice on memory is a well-established phenomenon. Despite the wealth of research on this *testing* effect, it is unclear whether the benefits of testing extend beyond the tested information to include memory for the context in which the memoranda were encountered. Three experiments examined the effect of testing on memory for context using a standard variant of a traditional item-context memory task, in which cue-target word pairs (the items) were presented on the computer screen in varying locations (the contexts). All experiments revealed an enhancement to memory for context following retrieval practice of the items, regardless of whether that retrieval took place in a neutral (Experiments 1 and 2) or in an interfering (Experiment 3) location. These results support the view that retrieval practice elicits retrieval of relatively comprehensive prior episodes, rather than of only semantic aspects of the prior episodes relevant to the practice cues.

Introduction

Retrieval of information from memory is a powerful means of enhancing long-term retention (Bjork, 1975). It is often more effective than additional study of the same information, a phenomenon called the *testing effect* (Roediger & Karpicke, 2006a; see Nunes & Karpicke, 2015; Rowland, 2014, for recent reviews). The benefits of testing have been demonstrated both in the lab and in classroom settings, using a variety of learning materials, including prose passages (e.g., Roediger & Karpicke, 2006b), single words (e.g., Hogan & Kintsch, 1971), paired associates (Carrier & Pashler, 1992), as well as nonverbal material (e.g., Wheeler & Roediger, 1992). Testing has even been promoted for wider use within educational settings as a means of enhancing, and not simply assessing, knowledge (Benjamin & Pashler, 2015; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler et al., 2007).

These important applications notwithstanding, there is lack of a consensus within the field as to what actually causes the benefits of retrieval practice on memory. Some theoretical positions include a prominent role for the episodic context of the original encoding, as well as of the retrieval practice event (e.g., Karpicke, and Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014), whereas others include no role for context (e.g., Carpenter, 2009) or are mute to its effects (e.g., Bjork, 1975; Kornell, Bjork, & Garcia, 2011). Within those theories that do allow a role for some kind of context, there are ones that attribute similarity between study and test circumstances as key (e.g., via transfer-appropriate processing, Landauer & Bjork, 1978; Roediger & Karpicke, 2006a, 2006b) and others in which variability across contexts is important (Karpicke et al., 2014; McDaniel & Masson, 1985). Clearly, experiments directly assessing the degree to which memory for contextual elements is enhanced, retarded, or unaffected by testing will be central to developing a thorough understanding of the causes of the testing effect. Here we report three experiments using traditional tests of context memory within a testing-effect paradigm, and demonstrate consistent enhancement to memory for context following retrieval practice. These benefits persist even when that retrieval practice introduces a context that would be expected to interfere with memory for the original encoding context.

A point that is highly relevant to the applied potential of testing effects is that the benefits of testing sometimes extend beyond the tested information itself to include conceptually related but nontested information presented in the same episode with the tested information (e.g., Butler, 2010; Carpenter, Pashler, & Vul, 2007; Chan, 2009, 2010; Chan, McDermott, & Roediger, 2006; but see Pan, Gopal, & Rickard, 2015). With semantically related materials it can be hard to tell whether such benefits reflect the incidental retrieval of untested aspects of the material or complex knock-on effects of enhancing memory for the tested material. For this reason, it is critical to evaluate this question using materials for which the untested elements are purely episodically related and devoid of larger meaning, thereby minimizing influences of semantic encoding and retrieval strategies. However, there are only a

* Corresponding author at: Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, United States.
E-mail address: makan2@illinois.edu (M. Akan).

https://doi.org/10.1016/j.jml.2018.07.003
Received 10 January 2018; Received in revised form 3 July 2018

handful of studies that examined whether the benefits of testing extend to contextual information under such circumstances, despite much research indicating a crucial role for context in episodic retrieval (Divis & Benjamin, 2014; Howard & Kahana, 2002; Jang & Huber, 2008; Lehman & Malmberg, 2013). Currently, there is no conclusive evidence as to whether the testing effect generalizes to memory for incidental source or context.

In one relevant study, Rowland and DeLosh (2014) found that the benefits of testing were not limited to untested items that were semantically related to the tested items, but also generalized to untested items that had no designated association with the tested items other than being presented as part of the same list. There is also evidence that participants are better able to identify the list membership of previously studied items following testing (Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010; Chan & McDermott, 2007; Verkoeijen, Tabbers, & Verhage, 2011), and that tested materials are more likely to elicit a *Remember* response in the remember/know paradigm (Jones & Roediger, 1995). However, neither of these data are dispositive. List membership is only a very coarse measure of context and is confounded with recency of exposure. And *remember* judgments may not accurately reveal retrieval of contextual information (Benjamin, 2005; Dunn, 2008).

To our knowledge, there is only one published study (Brewer et al., 2010) that examined whether testing leads to an enhancement in memory for context beyond episodic temporal information. In that task, participants studied two lists of words, and each word was presented in either a male or female voice. Both lists were followed by either retrieval practice (free recall) or a math distractor task. In the final test, participants indicated, for each studied item, either whether it had been presented in list 1 or in list 2, or whether the word had been spoken by a female or by a male voice. The results revealed an enhancement in memory for list membership but not for speaker gender. However, in another experiment, when participants were asked to also additionally indicate gender source information as they recalled each word during retrieval practice, testing also enhanced gender discrimination performance on the final test. These results would seem to indicate that temporal information is naturally accessed in the processes underlying cued recall, but that other contextual aspects of the original presentation are not unless the retrieval cue specifically promotes their involvement. Such a claim is buttressed by findings that temporal information is often automatically encoded, even under incidental learning conditions (Hintzman & Block, 1971; Proctor & Ambler, 1975). Yet, in contradiction with this claim, an unpublished thesis by Rowland (2011) reported that retrieval practice resulted in a small advantage in recalling which of the two possible colors a word was presented in (Experiment 1), and the order in which the individual words in semantically unrelated word pairs were presented (Experiment 2), even though the retrieval practice did not involve reporting either of these details.

In short, the few studies that directly examined memory for contextual information have not provided conclusive evidence. The present experiments used a time-uncorrelated contextual dimension and used more varied contextual information than previous work—the items could appear in one of eight possible locations on the screen. The spatial configuration of these screen locations was circular, as shown in Fig. 1a. Unlike a linear display, a circular arrangement mitigates against easy translation into a temporal code (Fischer-Baum & Benjamin, 2014; Hitch, 1974). Spatial information would seem to be more difficult to associate with words semantically than a gendered voice or temporal information (which allows for story-building and retrieval strategies that involve seriation), and may not be encoded automatically (Naveh-Benjamin, 1987, 1990). We directly compared testing with both a restudy condition and a control condition in which items did not receive any additional exposure.

Experiments 1A and 1B

Experiments 1A and 1B sought to test if the benefits of retrieval practice would extend to contextual details by having participants study word pairs presented in different locations on the screen. During the study session, words were presented in one of eight possible locations, and the review (either restudy or retrieval practice) occurred in the center of the screen. The stimuli in these experiments were low-association word pairs. Participants were asked to study the word pairs and were informed that there would be a later test in which they would be given the first word of the pair (the cue) and asked to provide the second word in the pair (the target). The only difference between Experiment 1A and Experiment 1B is the number of studied word pairs—96 and 48 respectively. The reason for reducing the number of word pairs for Experiment 1B was a concern over potential floor effects during the process of collecting data for Experiment 1A. Rather than starting over, we continued collecting data for Experiment 1, after reducing the number of items, until we achieved our planned sample size. This was determined to be n = 52 to achieve 80% power to detect an effect size of d = 0.40 for a paired-sample *t*-test.[1] As we did not know the effect size for retrieval practice on memory for context, we made a conservative estimate based on prior work on the effect of retrieval practice on item memory (see Rowland, 2014 for a meta-analysis).

Method

Participants

Twenty-nine undergraduate students from the University of Illinois at Urbana-Champaign (UIUC) participated in Experiments 1A and 1B, each, in partial fulfillment of a course requirement. Three participants in Experiment 1A had incomplete data and were excluded from analyses (two due to a failure to attend Day 2 of the experiment, and one due to computer difficulties). Four participants in Experiment 1B had incomplete data and were excluded from analysis (all due to a failure to attend Day 2 of the experiment). For all experiments reported in this paper except Experiment 2B, demographic information collected from participants was not connected to the particular experiments that they participated in. Here we provide the overall demographic profile of the subject pool from which the participants were drawn. Participants from this pool ranged from 18 to 35 years of age, and 91% of the participants were between the ages 18 and 21. Females constituted 63% of the subject pool and the percentage of native speakers was 78%.

Materials

Ninety-six weakly associated word pairs (cue to target association of 0.028–0.030) were selected from the University of South Florida Free Association Norms database (Nelson, McEvoy & Schreiber, 2004). For Experiment 1B, only 48 of the original 96 word pairs were used. We reduced the number of study pairs in Experiment 1B; the experiments were otherwise identical. The materials are included in the raw data files, which can be accessed online on our main project page at Open Science Framework (OSF; https://osf.io/bqr5f/).

Design

The experiment used a 3 (Review Type) × 2 (Type of Final Test) within-subject design. The three review types consisted of retrieval practice, restudy, and a control condition of no review. The final test was either a cued recall task that required retrieval of the target item given the cue item, or an 8-alternative-forced-choice (8-AFC) test on memory for the word location context. All conditions had an equal number of word pairs. Both review condition and test condition were manipulated between-item (i.e., no item was reviewed or tested more

[1] All power analyses were performed using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009).

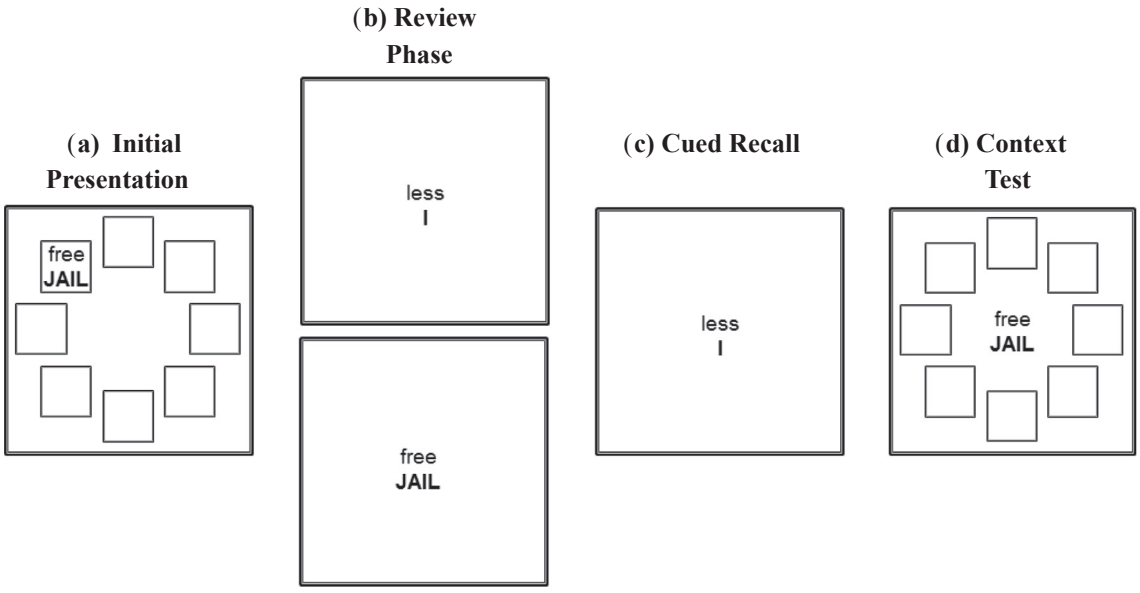

Fig. 1. Sample display of each phase. (a) Initial presentation: Eight boxes equidistant from the center of the screen with a word pair presented in one of the boxes. (b) Review Phase: Cue from word pair with a blinking cursor below presented in the center of the screen for retrieval practice trials (Experiments 1 and 2 only). Entire word pair was shown in the center of the screen for restudy trials. (c) Cued Recall (two days later): Cue from word pair with a blinking cursor below. (d) Context test: Word pair in the center of the screen surrounded by eight clickable boxes.

than once).

Procedure

Participants were told that they were going to be shown a series of word pairs and that they were to study them for a later test in which they would be prompted with the first word of a pair and asked to provide the second word from the pair. Participants were not told that they would later be tested on the location of the word pairs. The study phase presented the word pairs for 5 s each, with a 1 s inter-stimulus interval. The word pairs were presented on a computer screen in one of eight possible boxes, each equidistant from the center. For each subject, a quasi-random order was generated such that each box was used an equal number of times for each condition and no box was used twice in a row. After all the word pairs were presented, participants were given a 1-min distractor task in which they performed simple 1–2-digit addition.

During the review phase, participants performed retrieval practice on one-third of the word pairs, restudied one-third of the word pairs, and did not review the remaining one-third of the word pairs. For each participant, the word pairs were randomly assigned to conditions. For the word pairs that were assigned to the retrieval practice condition, a cued recall test was given in the center of the screen. Participants were instructed to guess if they did not know the answer, and they could not continue until at least three characters were typed. The retrieval practice was self-paced and did not include feedback. The word pairs assigned to the restudy condition were presented again for 5 s, also in the center of the screen. The retrieval practice and restudy conditions were randomly ordered with the constraint that no more than three of any one condition would appear in a row. The remaining third of the words comprised the control condition and were not revisited in this phase of the experiment. The Day 1 phase of the experiment ended after the review phase was completed; participants returned approximately 48 h later to complete the final test phase of the experiment.

For the final phase, two days later, participants were given a cued recall test in the center of the screen on one-half of the word pairs (an equal number of items from each of the three review conditions). Participants were instructed to guess if they did not know the answer, and they could not continue until at least three characters were typed. For the other half of the word pairs, participants were tested for context memory; a word pair was presented at the center of the screen along with the eight boxes from the study phase and they were asked to click on the box in which the word pair had originally been presented. The test types were blocked and the order of the test blocks was counterbalanced.

Results

The combined results of Experiments 1A and 1B are shown in Fig. 2. The two experiments exhibited the same pattern, and were combined for all analyses. Approximately half of the items were successfully recalled in the initial retrieval practice phase (M = .51, SD = .24). Final cued recall test performance was significantly higher in the retrieval practice condition (M = .46, SD = .28) than in the restudy condition

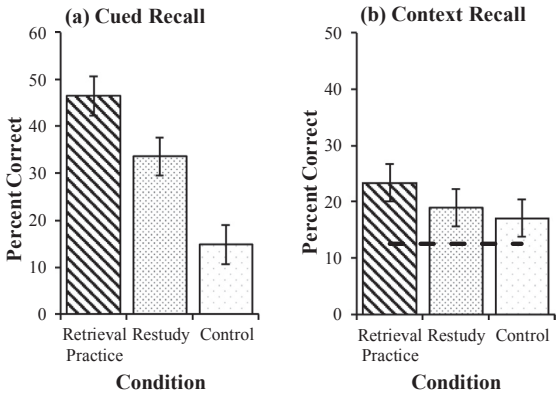

Fig. 2. Combined results of Experiments 1A and 1B. (a) Percentage of targets correctly produced on the final cued recall test in the retrieval practice, restudy and control conditions. (b) Percentage of contexts correctly selected in the retrieval practice, restudy and control conditions. The dashed line indicates chance performance if all options are equiprobable. The error bars here and on all subsequent graphs represent 95% confidence intervals and are based on the mean squared error for Subject × Condition interaction (see Equation 2 in Loftus & Masson, 1994). Because the confidence intervals are not based on the differences between pairwise conditions, they cannot be used to make statistical inferences about specific differences between any one condition and another.

(M = .34, SD = .23), $t(50)$ = 4.44, r = .68, p < .001, g = 0.62,[2] 95% CI [0.32, 0.92],[3] and it was significantly higher in the restudy condition than in the control condition (M = .15, SD = .14), $t(50)$ = 7.24, r = .58, p < .0001, g = 1.01, 95% CI [0.67, 1.35]. Performance on the context memory test was significantly higher in the retrieval practice condition (M = .23, SD = .15) than in the restudy condition (M = .18, SD = .13), $t(50)$ = 2.09, r = .40, p < .05, g = 0.29, 95% CI [0.01, 0.57] and it was numerically but not significantly higher in the retrieval practice relative to the control condition (M = .19, SD = .13), $t(50)$ = 1.62, r = 0.31, p = .11, g = 0.23, 95% CI [−0.05, 0.50]. We did not detect a difference across the restudy and control conditions in context memory performance (t < 1). Refer to Fig. 7 for a plot of all effect size estimates and their confidence intervals for the difference between retrieval practice and restudy conditions for all experiments.

Discussion

Memory for both items and their context was enhanced as a result of retrieval practice compared to both restudy or control (no review) conditions, indicating that the benefits of retrieval might extend beyond memory for tested information, and include contextual information associated with the tested items. However, while a medium-to-large effect size was observed for enhancement in item memory (g = 0.62), the effect size for context memory enhancement was small-to-medium (g = 0.29), and the benefit of retrieval practice over the control condition was not significant. These underwhelming results, along with the minor methodological difference across Experiments 1A and 1B, indicate that a high degree of certainty in the results would at this point be inadvisable. The remaining experiments provide an opportunity to replicate the benefit of retrieval practice on context memory.

We made one additional change in procedure in order to address a potential measurement confound in Experiment 1. Specifically, because context memory may be dependent on item memory, it is possible that the difference in memory for the word pairs (i.e., the traditional testing effect) is biasing our measure of the context memory. Perhaps, in order for people to successfully retrieve the context of an item, they must first successfully recall the word itself. If this were true, context memory test performance may be higher in the retrieval practice condition not because context memory is better, but because the memory for the word pair that serves as a cue for that context is better. What is needed is a way to conditionalize the context memory measures on item memory, but the between-item nature of the testing procedure in Experiment 1 precludes such an analysis. The remaining experiments address this concern by using a within-item manipulation of test type.

Experiment 2A and 2B

Experiment 2A and 2B were similar to Experiment 1, except that each word pair was subjected to both cued recall and context memory tests during the final test phase. This modification allowed us to examine context memory selectively for the words that were successfully recalled, thus removing the item confound from Experiment 1. Experiment 2B provides an exact replication of Experiment 2A with a larger sample size. Experiment 2A was run in parallel with Experiment 1B, and the sample size estimate for Experiment 2A was based on an effect size estimate of d = 0.40. However, because the effect actually turned out to be smaller, Experiment 2B is a replication with a planned sample size designed for the actually obtained effect size from Experiments 1, and 2A.

Method

Participants

Fifty-nine students from UIUC participated in Experiment 2A and 104 in Experiment 2B in partial fulfillment of a course requirement. In Experiment 2A, ten participants had incomplete data and were excluded from the analyses (three due to a failure to attend Day 2 of the experiment and seven due to computer problems). In Experiment 2B, 14 participants had incomplete data and were excluded from the analyses (nine due to a failure to attend Day 2 of the experiment, three due to computer problems and two due to experimenter error). Out of the 90 participants included in the analyses for Experiment 2B, 57 were females; mean age was 18.9 (SD = 1.58). Most of the sample (71 out of 90) were native English speakers.

Materials

The same 48 study word pairs were used as in Experiment 1B.

Procedure

The study phase and review phase were identical to Experiment 1B. In the final phase, each word pair was tested first using cued recall and then using forced-choice context memory test. The latter test was administered using the complete word pair as a prompt, regardless of the response on the first (cued recall) test. Immediately after recalling each target, participants were asked to indicate the location in which the cue-target pair was seen during the initial study phase. Everything else remained the same.

Results from Experiment 2A

The results of Experiment 2A are shown in Fig. 3. During the initial retrieval practice phase, 55% (SD = .25) of the targets were successfully recalled. As expected, final cued recall performance was significantly higher in the retrieval practice condition (M = .46, SD = .23), than in the restudy condition (M = .38, SD = .23), $t(48)$ = 3.01, r = .71, p < .01, g = 0.43, 95% CI [0.13, 0.72], which in turn was significantly higher than in the control condition (M = .17, SD = .13), $t(48)$ = 8.07, r = .59, p < .001, g = 1.15, 95% CI [0.78, 1.51]. Context memory was superior in the retrieval practice condition (M = .23, SD = .12) compared to both the restudy (M = .19, SD = .12), $t(48)$ = 2.09, r = .25, p < .05, g = 0.30, 95% CI [0.01, 0.58] and the control conditions (M = .17, SD = .10), $t(48)$ = 3.21, r = .24, p < .01, g = 0.46, 95% CI [0.16, 0.75]. Context memory did not differ across restudy and control conditions, t < 1.

In order to address the potential confound described at the end of Experiment 1, we also report a measure of context memory conditionalized upon successful item memory. Performance on the context memory test conditional upon successful recall during the final cued recall test was superior in the retrieval practice condition (M = .32, SD = .19) than in the restudy condition (M = .25, SD = .23), though this difference was not significant, $t(45)$ = 1.59, r = .003, p = .12, g = 0.23, 95% CI [−0.06, 0.53]. Three participants were omitted from this comparison because they did not successfully retrieve any words during the final cued recall test in either only the restudy condition (n = 1), only the retrieval practice condition (n = 1) or in both conditions (n = 1). There was no significant difference between performance in the restudy and the control conditions, t < 1. Five subjects

[2] We report Hedges' g for all of our effect size statistics, which corrects for the sample bias in estimating the population effect size (Hedges & Olkin, 1985). Even though the difference between Cohen's d and Hedges' g is small, especially for sample sizes greater than 20, it is preferable to report Hedges' g as meta-analytic work commonly relies on the latter (Lakens, 2013).

[3] Confidence intervals are for the effect size and they were estimated using the MBESS package in R. The estimation of confidence intervals for the effect size are based on the noncentral t-distribution when $\mu \neq \mu_0$. The noncentral t-distribution is a skewed distribution with two parameters; df (n − 1) and the noncentrality parameter (Δ). The observed t-value is used as an estimator of Δ. These two parameters, as well as the alpha, are input into the conf.limits.nct function in the MBESS package to get the t-values for the confidence interval limits (Kelley, 2007). We provide our R code for this computation in our main project page on OSF, https://osf.io/bqr5f/).

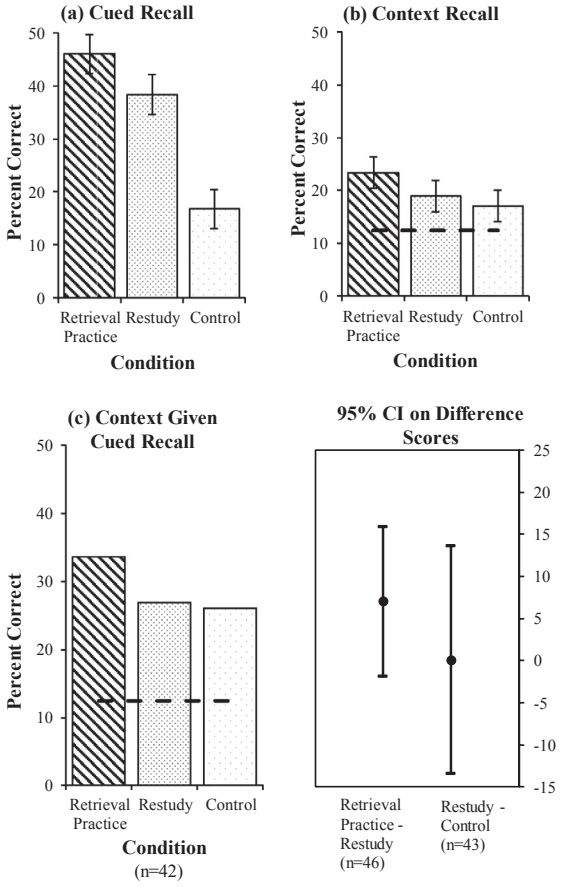

Fig. 3. Results of Experiment 2A. (a) Percentage of targets correctly produced on the final cued recall test in the retrieval practice, restudy and control conditions. (b) Percentage of contexts correctly selected in the retrieval practice, restudy and control conditions. (c) Percentage of contexts correctly recalled conditional upon correct cued recall of the target (left panel) and the 95% confidence interval for the mean difference between retrieval practice and restudy conditions, and restudy and control conditions (right panel). Conditionalized context recall scores were not computable when participants did not correctly recall any items; the corrected number of participants represented in each contrast is indicated in parentheses. The means displayed by the bar graphs on the left panel are based on participants who correctly recalled at least one item in each condition, whereas the mean differences in the right panel include all participants who recalled at least one item in two conditions that are compared. The dashed line indicates chance performance (12.5%) if all options are equiprobable.

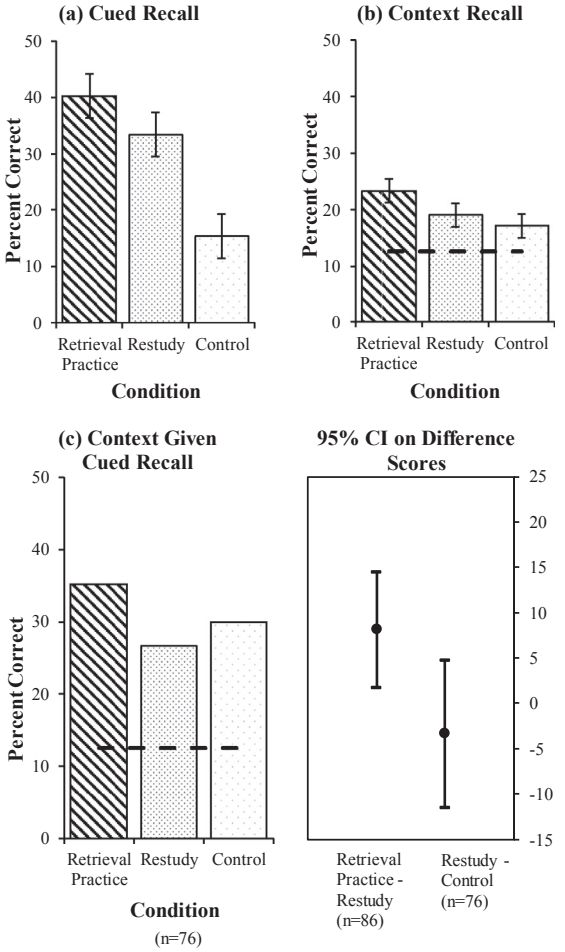

Fig. 4. Results of Experiment 2B. (a) Percentage of targets correctly produced on the final cued recall test in the retrieval practice, restudy and control conditions. (b) Percentage of contexts correctly selected in the retrieval practice, restudy and control conditions. (c) Percentage of contexts correctly selected conditional upon correct cued recall of the target. The dashed line indicates chance performance if all options are equiprobable. See the figure caption from Fig. 3 for additional information on participant inclusion in the lower portion of the figure.

were omitted from this comparison because they did not successfully retrieve any words only in the control condition (n = 4) or in both the restudy and the control conditions (n = 1).

Discussion of Experiment 2A

Experiment 2A again revealed a testing effect on memory for the word pairs. The context memory for all items (regardless of successful recall) replicated the results of Experiment 1: retrieval practice enhanced memory for context compared to restudy. The effect size for this difference was again small-to-medium (g = 0.30). Unlike in Experiment 1, a significant difference in context memory obtained across retrieval practice and control conditions. When context memory was conditionalized on item memory, the same pattern was evident, suggesting that the results from Experiment 1 results were not an artifact of the between-item testing procedure and dependence of context memory on item memory. However, the nonsignificant difference between retrieval practice and restudy when examining the subset of items for which the target was successfully retrieved continues to make it difficult to draw a firm conclusion about the effect of testing on context memory. Therefore, in Experiment 2B, we sought to replicate the context memory advantage in the retrieval practice condition with greater power. Because our observed effect size was g = 0.29 in Experiment 1 and g = 0.30 in Experiment 2A, we based our power computation on an effect size of 0.30 for Experiment 2B. For an effect size of 0.30, a sample size of 90 was planned to achieve 80% power for a paired *t*-test.

Results from Experiment 2B

The results of Experiment 2B are shown in Fig. 4. During the initial retrieval practice phase, 52% (*SD* = .21) of the targets in the testing condition were successfully recalled. Final cued recall performance was significantly higher in the retrieval practice condition (*M* = .40, *SD* = .21), than in the restudy condition (*M* = .33, *SD* = .23), *t*(89) = 3.62, r = .60, $p < .001$, *g* = 0.38, 95% CI [0.17, 0.59] which in turn was significantly higher than the control condition (*M* = .15, *SD* = .11), *t*(89) = 10.69, r = .55, $p < .001$, *g* = 1.13, 95% CI [0.86, 1.39]. Replicating Experiment 1 and 2A, context memory was superior in the retrieval practice condition (*M* = .24, *SD* = .17) compared to both the restudy (*M* = .19, *SD* = .11), *t*(89) = 3.06, r = .53, $p < .005$,

g = 0.32, 95% CI [0.11, 0.53] and the control conditions (M = .17, SD = .11), $t(89)$ = 4.06, r = .45, p < .001, g = 0.43, 95% CI [0.21, 0.64]. Context memory again did not differ across restudy and control conditions, $t(89)$ = 1.39, r = .35, p = .17. Unlike Experiment 2A, performance on the context memory test conditional upon successful recall during the final cued recall test was significantly higher in the retrieval practice condition (M = .35, SD = .25) than in the restudy condition (M = .27, SD = .25), $t(85)$ = 2.55, r = .33, p < .05, g = 0.28, 95% CI [0.06, 0.49]. Four participants were omitted from this comparison because they did not successfully retrieve any words during the final cued recall test in either only the restudy condition (n = 2), only the retrieval practice condition (n = 1), or both (n = 1). As in Experiment 2A, there was no significant difference between context memory given accurate item recall between the restudy and the control conditions, t < 1. Fourteen subjects were omitted from this comparison because they did not successfully retrieve any words in either only the control condition (n = 11) or both the restudy and control conditions (n = 3).

Discussion of Experiment 2B

Experiment 2B successfully replicated Experiment 2A in revealing a benefit of retrieval practice on both item and context memory. The greater power of Experiment 2B yielded significant results across all important comparisons, with effect sizes similar to what had been seen in the preceding experiments. The benefits of retrieval practice on context memory were apparent even when that measure was conditionalized on successful item recall. In other words, following successful recall of an item, the probability of recalling its context was higher if the item had received retrieval practice in an earlier session as opposed to restudied. This finding supports the idea that the benefits to context memory from retrieval practice are not solely due to downstream benefits to item memory. In Experiment 3, we sought to replicate the beneficial effect of retrieval practice on context memory under conditions that might naturally undermine that benefit: a retrieval practice phase that introduces interfering contextual information.

Experiment 3

One possible disadvantage of retrieval practice is that interfering memories for the review event could make it more difficult to distinguish context during the initial study from context during the review phase. Both Experiments 1 and 2 used the same "neutral" center position during the review phase and never used that position during the initial study phase. Therefore, there was little opportunity for confusion; participants knew that any recollection of a center position was either from a restudy or a retrieval practice event and that any other location was from the initial study event. Experiment 3 sought to produce interference during retrieval practice by having both restudy and retrieval practice occur in one of the boxes around the circle.

Experiment 3 was similar to Experiment 2, with the minor modification that during the review phase, items were presented in one of the boxes around the circle (as in the study phase), rather than in the center of the screen. No individual word pair was ever presented in the same location for both the initial presentation and the review phase. This procedure provides a more conservative test for the examination of benefits of retrieval practice on context memory relative to the prior experiments. The sample size for this experiment was based on the same power analyses conducted for Experiments 1 and 2A, as this experiment was also run prior to Experiment 2B.

Method

Participants

Fifty-four students from the UIUC participated in the experiment in partial fulfillment of a course requirement. Four participants were omitted from analysis due to a failure to attend Day 2.

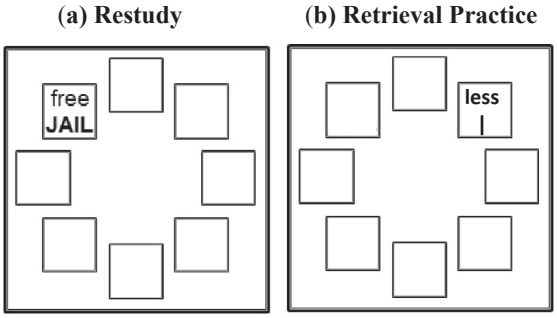

Fig. 5. Sample display for Experiment 3. The procedure was the same as in Experiments 1 and 2, but the review phase took place in one of the eight boxes.

Materials

The same materials were used as in Experiments 2A and 2B for all but the review phase. In the restudy phase, the word pairs were no longer in the center of the screen, as in prior experiments, but rather shown in one of the eight boxes used in the initial presentation phase (see Fig. 5a). In the retrieval practice condition, the prompt and blinking cursor were shown in one of the boxes (see Fig. 5b). For each word pair, the location during the review phase was determined randomly from the seven boxes left after eliminating the box that was used during the study phase. For both conditions, an individual word pair was never re-presented or tested in the same box in which it was originally presented.

Procedure

With the exception of the changes described above, the procedure was identical to Experiment 2.

Results

Results of Experiment 3 are shown in Fig. 6. During the retrieval practice phase, 54% (SD = .24) of the targets were successfully recalled. Final cued recall performance was significantly higher in the retrieval practice condition (M = .42, SD = .22) than in the restudy condition (M = .33, SD = .18), $t(49)$ = 4.62, r = .78, p < .001, g = 0.65, 95% CI [0.34, 0.96], which in turn was higher than the control condition (M = .15, SD = .12), $t(49)$ = 8.67, r = .61, p < .001, g = 1.22, 95% CI [0.85, 1.59]. We again found that context memory was superior in the retrieval practice condition (M = .22, SD = .11), relative to restudy (M = .17, SD = .09), $t(49)$ = 2.97, r = .14, p < .01, g = 0.42, 95% CI [0.13, 0.71] and in the restudy relative to the control condition (M = .13, SD = .09), $t(49)$ = 2.58, r = .33, p < .05, g = 0.37, 95% CI [0.08, 0.65], in which performance was not detectably above chance levels, t < 1. The performance on the context memory test conditional upon successful recall of the target words was significantly higher in the retrieval practice condition (M = .31, SD = .23) than the restudy condition (M = .19, SD = .18), $t(47)$ = 2.68, r = .001, p < .01, g = 0.39, 95% CI [0.09, 0.68]. Two participants were omitted from this comparison because they did not successfully retrieve any word in either only the restudy condition (n = 1) or only the retrieval practice condition (n = 1). The performance on the context memory test conditional upon successful recall of the target words did not differ across restudy and control conditions, t < 1. Eight subjects were omitted from this comparison because they did not successfully retrieve any word in either only the restudy condition (n = 1) or only the control condition (n = 7).

Discussion

Experiment 3 again successfully demonstrated a benefit of testing

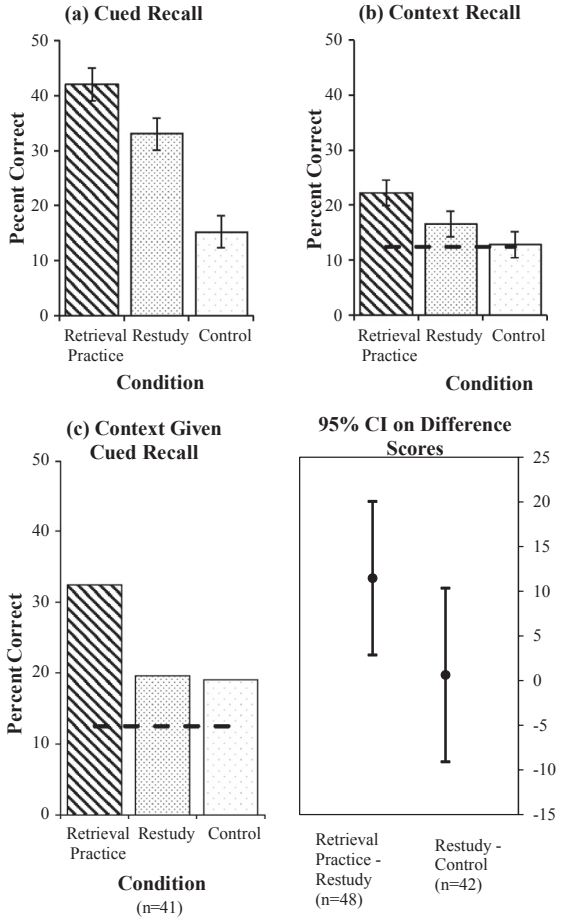

Fig. 6. Results of Experiment 3. (a) Percentage of targets correctly produced on the cued recall test in the retrieval practice, restudy and control conditions. (b) Percentage of contexts correctly selected in the retrieval practice, restudy and control conditions. (c) Percentage of contexts correctly selected conditional upon correct cued recall of the target. Dashed line indicates chance performance if all options are equiprobable. See the figure caption from Fig. 3 for additional information on participant inclusion in the lower portion of the figure.

for both the memory for the word pairs and the memory for the context in which the pairs were presented. Retrieval practice enhanced context memory even though the review phase took place in an interfering context. This result indicates that testing does not make it more difficult to distinguish context during the original study from the context in the review phase. Furthermore, the benefit of testing to context memory was significant even when considering only trials in which the target was successfully recalled. This result indicates even more strongly that there is a benefit of testing to context memory beyond improved memory for the word pairs.

It is worth noting that, even though the face validity of our interference manipulation is strong, we did not include in this experiment an assessment of interference, and consequently do not provide definitive evidence for having successfully created interference. We did observe a slight reduction in context memory in this experiment compared to previous experiments which did not involve interfering contexts. However, this does not constitute strong evidence for interference due to potential differences in the participant samples across experiments.

General discussion

In a series of four experiments, we consistently found that retrieval practice enhanced memory both for cued recall of target items as well as memory for their context (i.e., location on the screen) relative to restudy and control conditions. Context memory was enhanced even when the retrieval practice phase introduced interference by presenting the word pairs in locations that differed from the locations in the initial study phase. To our knowledge, this is the first study to show a beneficial effect of retrieval practice on memory for contextual information that is not directly tied with the tested material semantically or temporally.

One concern of designs like the ones employed here is that the self-paced nature of the retrieval practice allows the potential for confounds arising from overall time allocated to processing the items from different conditions. We examined whether the time allocated to reviewing the items differed between the retrieval practice and restudy conditions. Averaged across all experiments, there was no detectable difference between the time spent during retrieval (M = 5.15, SD = 2.08) and restudy (5 s) during the review phase. Thus, the time spent reviewing items is not likely to be an adequate explanation for the benefits of testing on item or context recall.

To assess the big picture across the experiments presented here, we computed the effects collapsed across experiments and calculated Bayes factors for the effect of retrieval practice compared to restudy for context memory, as well as for item memory, from all experiments (N = 240). Bayes factors provide a ratio of the amount of evidence in favor of one prespecified hypothesis over another, and are commonly used in conjunction with null-hypothesis testing to express relative evidence for an alternative hypothesis over a null hypothesis (BF_{10}). Here all Bayes factors were computed using the Jeffreys-Zellner-Siow (JZS) prior advocated by Rouder, Speckman, Sun, Morey, and Iverson (2009), with the r scale for the Cauchy prior set to 0.707. As proposed by Jeffreys (1961), a BF_{10} of 1-to-3 was interpreted as little to no evidence in support of the alternative hypothesis, a BF_{10} of 3-to-10 as substantial evidence, and a BF_{10} greater than 10 as very strong evidence. The aggregate analyses revealed an enhancement in final cued recall in the retrieval practice condition (M = .43, SD = .23) compared to the restudy condition (M = .34, SD = .21), $t(239) = 7.60$, r = .67, $p < .001$, $BF_{10} = 10 \times 10^7$. The effect size can be considered of medium-to-large magnitude (g = 0.49, 95% CI [0.36, 0.62]), and is slightly smaller than the effect sizes (g = 0.57–0.61) reported in a meta-analysis of the testing effect (Rowland, 2014). Memory for context was also higher in the retrieval practice (M = .23, SD = .14) than the restudy condition (M = .18, SD = .11), $t(239) = 4.11$, r = .39, $p < .001$, $BF_{10} = 230.40$, with a small-to-medium effect size (g = 0.33, 95% CI [0.20, 0.46]). For both item memory and context memory, Bayes Factors indicated very strong evidence for the beneficial effect of retrieval practice over restudy.

A further analysis pooled the data only from the experiments in which all items received both item and context memory tests at the final retrieval phase (Experiments 2A, 2B and 3; N = 189), and revealed that performance on the context memory test conditionalized upon successful recall on the final test was significantly higher in the retrieval practice condition (M = .32, SD = .23) than the restudy condition (M = .24, SD = .23), $t(179) = 4.01$, r = .20, $p < .001$, $BF_{10} = 163.06$), with a small-to-medium effect size (g = 0.30, 95% CI [0.15, 0.45]). The results indicate that the superior memory for context appeared even when controlling for successful item memory, and thus cannot be a consequence of superior memory for the word-pair cue.

Memory for context was not better in the restudy condition (M = .18, SD = .11) relative to the control condition (M = .17, SD = .11), even when combining all three experiments, $t(239) = 1.84$, p = .07—a contrast that had very high power (.996) to find a small-to-medium effect size of 0.30. The Bayes factor computed for this comparison revealed an estimate of $BF_{01} = 2.63$, indicating equivocal evidence. This null result is unlikely to be a floor effect since memory for context was significantly above chance for both restudy and control conditions, $t(239) = 8.05$, $p < .001$ and $t(239) = 5.92$, $p < .001$, respectively. It thus seems that not all methods of study improve memory

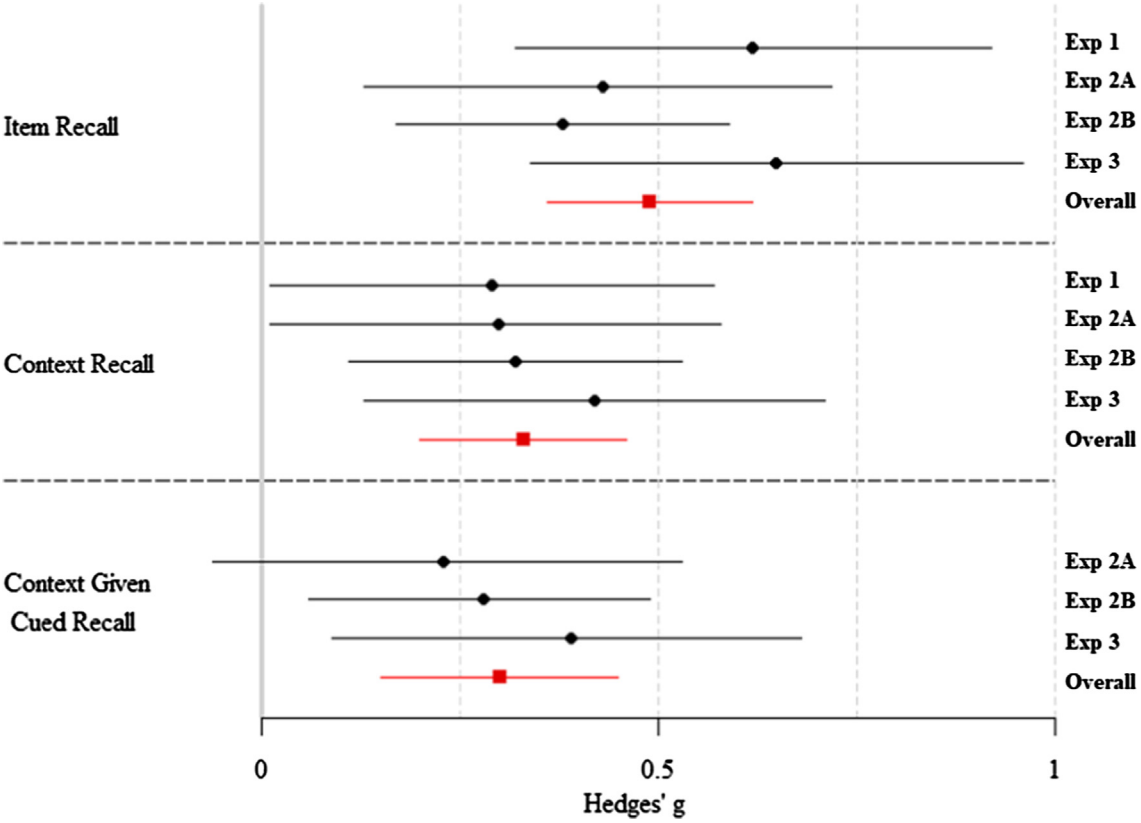

Fig. 7. Effect size estimates (Hedges' g) and their 95% confidence interval for the difference between retrieval practice and restudy conditions for item, context, and context given item recall across Experiments 1, 2A, 2B, 3, and pooled data from all experiments.

for context, even if they improve memory for the target. Even though restudying the items benefited memory for the individual items, it did not benefit memory for the contextual details of the episode in which the items were encountered. This is presumably because restudying does not compel retrieval of the relevant prior episode.

Our results indicate that retrieval not only enhances memory for the retrieved item itself but also for the incidental contextual details that accompanied the item during its initial encoding. Contrary to Brewer et al. (2010), we found that such memory enhancement for contextual details occurred even though the retrieval practice phase in the current set of experiments never required the retrieval of these details. We found these benefits even when examining only those trials in which the word pair was successfully recalled on the final test, indicating that the improved memory for context was not a consequence of improved memory for the items. Moreover, memory for context was superior even in a condition where the context of the retrieval practice was designed to interfere with the context memory of the prior study episode. Taken together, these results suggest that the benefits of retrieval extend beyond memory for the involved memoranda. Retrieval appears to enhance our memories for the original encoding *episode*, which includes aspects that are not directly tied to the material being tested.

These findings are in line with the prominent view of episodic retrieval as involving reinstatement of the original encoding context (e.g., Johnson, Hashtroudi, & Lindsay, 1993; Tulving, 2002). In the current experiment, retrieval practice during the review phase might have encouraged participants to reinstate the original encoding context by retrieving the original location of the cue-target pair, thereby enhancing its retention. Our findings are also in line with findings suggesting that active retrieval promotes binding of the retrieved content with other episodic elements (Bridge & Voss, 2014, 2015; Shimamura & Wickens, 2009; Shimamura, 2011).

In light of these findings, it can be speculated that retrieval of target words during retrieval practice promoted additional binding of cue-target pairs with their retrieved locations from the initial encoding phase. If so, during the final criterion test, the items that underwent retrieval practice acted as stronger cues for the retrieval of location information. This "binding" hypothesis can be tested by comparing the degree to which contextual cues support item memory across retrieval practice and restudy conditions. If testing leads to a greater degree of integration between the items and their contextual details, one might expect contextual cues to act as stronger cues for tested items. Our procedure does not provide results that speak to this hypothesis, however.

In summary, the current experiments clearly show enhancements to memory for contextual details for items that have undergone retrieval practice. As such, it suggests that the enhancement of memory for contextual details that go beyond the broader episodic/temporal context might be a critical element that theories of testing need to consider.

Author note

Melisa Akan, Department of Psychology, University of Illinois at Urbana-Champaign.

Sarah E. Stanley, Department of Psychology, University of Illinois at Urbana-Champaign.

Aaron S. Benjamin, Department of Psychology, University of Illinois at Urbana-Champaign.

Sarah E. Stanley is now at Wolfram Research.

The authors would like to thank the members of the Human Memory and Cognition Lab for their feedback throughout the execution of this work, and Opal Harshe for her assistance with data collection.

References

Benjamin, A. S. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a

caution about purportedly nonparametric measures. *Memory & Cognition, 33*, 261–269.
Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: A perspective from Cognitive Psychology. *Policy Insights from the Behavioral and Brain Sciences, 2*, 13–23.
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.
Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory, 18*, 385–393.
Bridge, D. J., & Voss, J. L. (2014). Active retrieval facilitates across-episode binding by modulating the content of memory. *Neuropsychologia, 63*, 154–164.
Bridge, D. J., & Voss, J. L. (2015). Binding among select episodic elements is altered via active short-term retrieval. *Learning & Memory, 22*, 360–363.
Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118–1133.
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., Pashler, H., & Vul, E. (2007). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*, 826–830.
Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633–642.
Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153–170.
Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*, 49–57.
Chan, J. C., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431.
Chan, J. C., 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*, 553–571.
Divis, K. M., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition, 42*, 1049–1062.
Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.
Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review, 115*, 426–446.
Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G∗Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.
Fischer-Baum, S., & Benjamin, A. S. (2014). Time, space, and memory for order. *Psychonomic Bulletin & Review, 21*(5), 1263–1271.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.
Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology, 88*, 297–306.
Hitch, G. J. (1974). Short-term memory for spatial and temporal information. *Quarterly Journal of Experimental Psychology, 26*, 503–513.
Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*, 562–567.
Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269–299.
Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 112–127.
Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.
Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.
Jones, T. C., & Roediger, H. L., III (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology, 7*, 65–80.
Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.). *Psychology of Learning and Motivation* (pp. 237–284). San Diego, CA: Elsevier Academic Press.
Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1–24.
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.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.
Landauer, T. K., & Bjork, R. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.). *Practical aspects of memory* (pp. 625–632). New York: Academic Press.
Lehman, M., & Malmberg, K. J. (2013). A buffer model of encoding and temporal correlations in retrieval. *Psychological Review, 120*, 155–189.
Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1787–1794.
Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476–490.
McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 11*, 371–385.
Naveh-Benjamin, M. (1987). Coding of spatial location information: An automatic process? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 595–605.
Naveh-Benjamin, M. (1990). Coding of temporal order information: An automatic process? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 117–126.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*, 402–407.
Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: Research at the interface between cognitive science and education. In R. A. Scott, & S. M. Kosslyn (Eds.). *Emerging trends in the social and behavioral sciences* (pp. 1–16). John Wiley Sons, Inc.
Pan, S. C., Gopal, A., & Rickard, T. C. (2015). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 108*, 1–13.
Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007–2004).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Research.
Proctor, R. W., & Ambler, B. A. (1975). Effects of rehearsal strategy on memory for spacing and frequency. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 640–647.
Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.
Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.
Rowland, C. A. (2011). *Testing effects in context memory (Unpublished master's thesis).* Fort Collins: 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., & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review, 21*, 1516–1523.
Shimamura, A. P. (2011). Episodic retrieval and the cortical binding of relational activity. *Cognitive, Affective, & Behavioral Neuroscience, 11*, 277–291.
Shimamura, A. P., & Wickens, T. D. (2009). Superadditive memory strength for item and source recognition: The role of hierarchical relational binding in the medial temporal lobe. *Psychological Review, 116*, 1–19.
Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1–25.
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.
Wheeler, M. A., & Roediger, H. L., III (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240–245.