Mem Cogn (2014) 42:1049–1062
DOI 10.3758/s13421-014-0425-y

Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning

Kristin M. Divis · Aaron S. Benjamin

Published online: 11 June 2014

Abstract In recent work, retrieval has been shown to enhance memory for events following that retrieval. In this set of experiments, we examined the effects of interleaved *semantic* retrieval on both previous and future learning within a multilist learning paradigm. Interleaved retrieval led to enhanced memory for lists learned following retrieval. In contrast, memory was impaired for lists learned prior to retrieval (Experiment 1). These results are consistent with recent work in multilist learning, directed forgetting, and list-before-last retrieval, all of which indicate a crucial role for retrieval in enhancing mental list segregation. This pattern of results follows clearly from a theoretical perspective in which retrieval drives internal contextual change and in which contextual overlap between study and test promotes better memory. Consistent with that perspective, a 15-min delay before the final test eliminated both effects (Experiment 2). Experiment 2 replicated the results of Experiment 1 with materials and assessments more appropriate for educational settings: Interleaved semantic retrieval led learners to be more able to answer questions correctly about texts studied after a retrieval event but less able to do so for texts studied earlier.

Keywords Memory · Context change · Interference

Testing has many beneficial effects on memory. It is well established that testing previously learned material enhances long-term memory for that tested material (e.g., the *testing effect*; for a review, see Roediger and Karpicke 2006). Retrieval events can also influence learning by affecting proactive interference (PI; Pastötter, Schicker, Niedernhuber and Bäuml 2011; Szpunar, McDermott and Roediger 2008) and retroactive interference (RI; Jang and Huber 2008), as we will review in more detail below.

The goals of the present set of experiments are to evaluate the effects of retrieval on both future and prior learning within a single experimental paradigm and to more precisely determine the origin of the costs and benefits of retrieval. The working hypothesis presented here, developed from prior work by Sahakyan and Kelley (2002) and Jang and Huber (2008), is that retrieval events lead to greater internal context change and, thus, to greater contextual segregation between events prior to and after the retrieval event. This segregation has the potential to improve retention (by decreasing the interference between competing events) and also to impair retention (by creating a greater disparity between the context present during encoding and the one present during the eventual criterion test). In Experiment 1, we examine the effects of interleaved semantic retrieval on both early and later learning of simple material when a final test immediately follows the study session. Experiment 2 replicates the results of Experiment 1 with more complex text materials. Experiment 3 extends the findings of Experiment 1 with a delayed final test.

The effects of testing on future learning

Although testing is mostly known for its large effect of enhancing memory for the tested material itself, testing also has an influence on both future and prior learning. We will first focus on the effect retrieval has on later learning: overall enhanced performance and a reduction in PI.

In one of the first studies examining how testing influences future learning, Tulving and Watkins (1974) explored the consequences of retrieval within an AB–AC interference paradigm. Having an immediate test following study of the AB list led to superior memory for the yet-to-be-learned AC items than when no test was given after the AB list. One

K. M. Divis (✉) · A. S. Benjamin
Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, IL 61820, USA
e-mail: divis1@illinois.edu

interpretation of this effect is that second-list learning was impaired when the first list was not tested before studying the second list. Other research supported this interpretation by demonstrating that words from untested lists were also more likely to intrude into recall on later tests than words from tested lists (Darley and Murdock 1971). These results indicate that testing may reduce the buildup of PI, a claim supported in later work by Szpunar et al. (2008): When participants were given either extended study sessions or interpolated tests while studying lists of words, those in the interpolated test condition showed a marked reduction in PI across lists. The results of additional experiments indicated that PI builds over time, becoming greater as more studied lists remained untested before recall of the final list.

Interestingly, the reduction in PI that is apparent following retrieval is not limited to conditions in which the retrieved material is from the preceding (and thus potentially interfering) list. Pastötter et al. (2011) examined the effects of alternate forms of retrieval. Participants completed one of five tasks between studying five different lists of words: a distractor task (counting backward by 3s), restudying the immediately preceding list, free recall of the immediately preceding list, a 2-back short-term memory task, or a semantic retrieval task (generating examples from a category like “sports”). Immediate recall and final recall of the last list were enhanced for those in the three retrieval conditions (free recall, *n*-back, and semantic retrieval), as compared with those in the distractor and restudy conditions, suggesting that the process of retrieval itself (and not just retrieval of the potentially interfering items) enhances future learning. Such a result rules out the previously plausible hypothesis that the reduction in PI is driven by better source memory (due to enhanced memory for the tested items). The fact that the benefit occurs (and is of approximately equal magnitude) when the interleaved retrieval does not actually test memory for (and thereby enhance) any of the previously studied material suggests a different basis for the reduction in PI.

The effects of retrieval on past learning

Retrieval has a beneficial influence on later learning (appearing to derive from a reduction in PI following the event); however, the influence of retrieval on *prior* learning is not as uniformly positive. The majority of studies addressing the effects of retrieval on prior learning are studies of the “testing effect,” in which the retrieved material is the prior list itself. In that case, it is quite clear that retrieval provides a substantial and long-lasting benefit to memory for that prior list (for a review of the testing effect, see Roediger and Karpicke 2006). However, there are also hints in the less well known *list-before-last* paradigm (Shiffrin 1970) that the act of retrieval fundamentally affects even previously learned material that is *not* retrieved. This type of retrieval appears to enhance list isolation but lead to overall lower performance on material learned prior to the retrieval event.

Shiffrin (1970) asked participants to recall words from the list immediately prior to the most recently studied list (e.g., if they studied list 1 followed by list 2, they were asked to recall words from list 1). Although the length of the to-be-recalled list influenced memory (the *list length effect*; Murdock 1960; Roberts 1972), the length of the intervening list did not. This result suggests that participants were able to effectively exclude the most recent event from consideration and thereby avoid RI (which would be expected to increase with the length of the interfering list). Later work showed that this effect obtains only when a retrieval event occurs between the two lists; in the absence of retrieval, shorter intervening lists lead to better performance on the tested list than do longer intervening lists (Jang and Huber 2008). Like the results considered in the previous section, this finding suggests that retrieval decreases the degree to which the lists compete with one another at retrieval. Jang and Huber hypothesized that recall created a significant context change between the lists and that this change caused the intervening list to interfere less with retrieval of the target list. They created a model of the effects of context similarity and context retrieval within a multilist learning environment to support their hypothesis. Importantly for the present set of experiments, Jang and Huber also examined the influence of a semantic lexical completion task and found evidence for list isolation (replicating the free recall results).[1] Notably, overall performance for the prior list was *reduced* in the retrieval conditions, as compared with the control. Further research utilizing the list-before-last paradigm also found that intervening retrieval events (as opposed to math problems) led to reduced intrusions but also, overall, reduced performance, as compared with a control condition (Sahakyan and Hendricks 2012; Sahakyan and Smith 2013). Taken together, the results here indicate that the consequences of retrieval are twofold: Although retrieval often benefits memory by reducing interlist interference, it also reduces access to material learned earlier.

The context change hypothesis and directed forgetting

Retrieval appears to lead to the opposite effects of enhancing performance on future items but hindering performance on prior items. The context change hypothesis of the effects of

[1] Jang and Huber (2008) also implemented a short-term memory *n*-back task that did not lead to list isolation (unlike other types of retrieval). In contrast, Pastötter et al. (2011) also used a short-term memory *n*-back task that enhanced future learning (like other types of retrieval). It is worth noting that retrieval from short-term memory may lead to different results than retrieval from long-term memory. However, the focus of this article is on semantic retrieval (a form of long-term retrieval).

retrieval provides a way of understanding both of these phenomena.

The context change hypothesis posits that during encoding, fluctuating contextual cues are bound to an internal context (Estes 1955; Mensink and Raaijmakers 1988). Context is important because performance is enhanced when the context at time of retrieval resembles the study context (e.g., Estes 1955; Kahana and Howard 2005; Tulving and Thomson 1973). Here, we assume that a retrieval event serves to enhance context fluctuation, thus binding subsequently encoded information to a more different internal context than would otherwise be the case. Similar hypotheses have been proposed to explain the underlying effects of retrieval in multilist learning (e.g., Jang and Huber 2008; Pastötter et al. 2011; Sahakyan and Hendricks 2012), as well as the *directed forgetting* effect (e.g., Sahakyan and Kelley 2002). Below, we briefly review the directed-forgetting literature, since that is where the influence of context change has been most closely studied.

In a typical *list-method directed-forgetting* study, participants study list 1 and are instructed to either remember or forget that list (usually between participants). They then study list 2 and are eventually given recall tests for one or both lists. Participants in the *forget* condition recall fewer items from list 1 but more items from list 2 than do participants in the *remember* condition (Reitman, Malin, Bjork and Higman 1973), suggesting that intentional forgetting of list 1 reduces PI (Bjork and Bjork 1996).

However, effects similar to those of directed forgetting can be achieved by means other than intentionally trying to forget. Manipulations that mimic directed-forgetting effects include imagining being invisible (Sahakyan and Kelley 2002), imagining walking through one's childhood home (Sahakyan and Kelley 2002), or imagining a vacation (Delaney, Sahakyan, Kelley and Zimmerman 2010). In fact, effects of greater magnitude were seen when the imagined event was further away, in either time or space (Delaney et al. 2010). Tasks that do *not* mimic directed-forgetting effects include solving math problems (Sahakyan and Kelley 2002), number searches (Mulji and Bodner 2010), speeded reading (Delaney et al. 2010), and counting tasks (Pastötter and Bäuml 2007; Pastötter et al. 2011; Sahakyan, Delaney and Goodmon 2008). Most of the tasks that mimicked directed-forgetting effects involved mental context change and retrieval of some sort (and imagining events may involve many of the same processes as retrieving them; e.g., Schacter, Addis and Buckner 2008). While the consequences of and mechanisms underlying a directed-forgetting cue are not identical to those of semantic retrieval, context change underlies both tasks. Insight into context change in the directed-forgetting paradigm informs predictions for context change in semantic retrieval.

Traditionally, measures of context change have suffered from a circular pattern of reasoning where context change predicts memory performance and is shown to exist because of changes in memory performance (e.g., Sahakyan and Smith 2013). However, our theoretical understanding of context change has allowed for accurate predictions in novel paradigms. Additionally, work by Sahakyan and Smith (2013) has recently shown that the influence of context change generalizes from memory performance to perceived estimation of time, thus potentially allowing future experiments on learning to examine context change independently of tests of memory.

A hypothesis of the effects of retrieval on context

In the present experiments, we take the context change hypothesis as our basis for understanding the varied effects of retrieval on multilist learning, specifically addressing both prior and future learning, as well as the effect of test delay. We propose that retrieval serves to enhance context fluctuation, causing pre- and post-retrieval items to have a greater contextual disparity than would have occurred without the retrieval event. Having different contextual cues tied to different lists promotes list isolation and reduces interference. Greater context change should lead to (1) *enhanced* performance on later lists because of a reduction in PI (e.g., Pastötter et al. 2011; Szpunar et al. 2008) and (2) overall *reduced* performance on earlier lists. The latter effect occurs because, although RI is reduced, the enhanced context fluctuation renders the earlier list context more dissimilar to the test context, and thus it is more difficult to reinstate the earlier context (e.g., Jang and Huber 2008; Sahakyan and Hendricks 2012). In other words, when multiple retrieval events occur between the earlier list and the test, the earlier list is further back in the contextual stream, relative to the criterion test (and thus more difficult to access). In addition, our hypothesis predicts that a delay prior to test should reduce the magnitude of both of these effects because the large delay-induced difference in context between test and original study contexts will overpower the relative difference within the study contexts caused by retrieval. See Appendix 1 for a simple mathematical model that supports these predictions (and is fitted to the data from Experiments 1 and 3).

Present experiments

The goal of the present study is to further understand the effects of retrieval on multilist learning. In order to achieve an uncontaminated look at both initial-list and last-list effects, we used a semantic retrieval task similar to that used by Pastötter et al. (2011), rather than the more commonly used episodic retrieval task (recalling either the immediately preceding list [e.g., Szpunar et al. 2008] or the list-before-last [e.g., Jang and Huber 2008; Shiffrin 1970]). Using the more classic method of episodic retrieval of previously studied material hinders the evaluation of the effect of retrieval itself on prior learning. Episodic retrieval of earlier material either enhances the material in question by retrieving it (i.e., the testing effect) or enhances potentially interfering material (as in the list-before-last paradigm). Semantic retrieval has been shown to have effects similar to that of episodic retrieval when its influence on prior learning (Jang and Huber 2008) and future learning (Pastötter et al. 2011) is evaluated, while avoiding the potential confound of enhancing memory for previously studied and potentially competing items. Thus, semantic retrieval is the better method with which to examine the effects of retrieval itself without the detrimental confounds introduced by episodic retrieval. Experiment 1 examined the effects of interleaved retrieval on initial-list and last-list performance with an immediate free recall test, Experiment 2 extended the results of Experiment 1 to more educationally relevant text materials, and Experiment 3 introduced a 15-min delay before the final test (using materials similar to Experiment 1).

Experiment 1

Method

Participants

Eighty-six undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment for course credit. Data from 7 participants were dropped because they did not follow the instructions (they recalled words from the previous list, instead of generating exemplars from the given category).

Design

Type of intervening task (unrelated semantic retrieval or control) and testing order of the studied lists were both manipulated between participants. Memory performance was measured as the proportion of words correctly free recalled.

Materials

Fifty words (average word length = 5.14 letters, *SD* = 1.46) were drawn from the University of South Florida Free Association Norms (Nelson, McElvoy and Schreiber 1998). A random subset of 10 words was used for each of five lists. Only words unrelated to each other were used (and none were members of the semantic categories used in the intervening task). No words were repeated in the experiment.

Procedure

Participants worked individually in a small room. PC-style computers programmed using MATLAB with the Psychophysics Toolbox extensions (Brainard 1997; Pelli 1997) were used to present stimuli and record responses. Prior to the initial study phase, participants were presented with a set of instructions informing them that they would be studying lists of words for a later free recall test. An example was also given. All words were presented for 4 s each, with an interstimulus interval of 500 ms. The procedure was adapted from Pastötter et al. (2011). Participants studied five lists of words, each separated by an intervening task. The distractor task consisted of counting backward by 3s from a randomly generated three-digit number for 30 s. Participants in the distractor condition were given three sets of these counting tasks (for a total of 90 s spent on the task). Participants in the retrieval condition first had one 30 s set of counting backward by 3s (like the distractor condition)[2] and then spent 60 s on a semantic retrieval task. They were given 60 s to type in as many exemplars as they could from a given category (four-legged animals, sports, vegetables, or professions, randomly ordered across the intervening tasks). Instructions were given before each task. After studying the final list (list 5), all participants were given two more sets of counting backward by 3s (for a total of 60 s). Half of the participants were then given a free recall test on list 1, and half were tested on list 5. For completeness, and to avoid the appearance of deception in the initial study instructions, this test was followed by the test for the other list. However, because the theoretical stance developed here makes no clear predictions about how this first episodic test should affect performance on the second, we restrict our analysis here to performance on the first test. All responses were typed into the computer, and no response time limit was imposed.

Results and discussion

Significance levels for all statistical tests were set at an $\alpha < .05$ level. In order to account for error terms that were not

[2] For consistency with previous work examining both episodic (e.g., Szpunar et al. 2008) and semantic (Pastötter et al. 2011) retrieval, we included a brief session of the control math task immediately prior to the retrieval task. The value of this methodology is that it should clear primary memory (e.g.., Glanzer and Cunitz 1966) and allow for a similar transition out of the study state of the previous test into the intervening task. It also makes the experience of the 60 s counting task that follows the last list a roughly equally distinctive event for both groups. The cost of this choice is that the semantic retrieval condition includes two tasks and the control condition only a single one. However, prior work (Pastötter et al. 2011) has shown that an intervening task of counting followed by restudy led to similar effects as the single task of only counting, both of which were different from counting plus retrieval tasks. This set of results suggests that this difference between conditions is unlikely to meaningfully affect the magnitude of induced context change between conditions.

distributed normally and be able to include both participant and item variability, mixed effects models were used, rather than the standard analysis of variance statistical tests (for further information on these methods, see Baayen, Davidson and Bates 2008, and Jaeger 2008). All models were fitted via Laplace estimation with the lme4 package (Bates, Maechler and Dai 2011) in R software (R Development Core Team 2008). As was expected, testing order significantly influenced the results, so only the results of the first test are reported here (see Appendix 2 for results and discussion of the second test).

The best-fit model included the fixed effects of list (1 or 5) and intervening task (retrieval or distractor) along with random intercepts for participants. Adding items as a random effect did not reliably improve the fit of the model, $\chi^2 = 0.990$, $p = .320$, so it was not included in the final model. Mean proportion correct are shown in Fig. 1a. The model indicated no overall effect of task; however, there was a main effect of list, $z = 3.629$, $p < .001$, with higher overall performance in list 5 than in list 1. Furthermore, a significant two-way interaction, $z = 3.164$, $p = .0016$, revealed higher performance in list 1 for the distractor condition, as compared with the retrieval condition, but higher performance in list 5 for the retrieval condition, as compared with the distractor condition. This pattern of results was supported by reliable simple effects of task in both list 1, $z = 2.002$, $p = .0453$, and list 5, $z = 2.500$, $p = .0124$.

Interleaved semantic retrieval events (as compared with a nonmnemonic distractor task) led to overall *reduced* performance for list 1 recall but *enhanced* performance for list 5 recall in tests immediately following the study session. This pattern of results follows from the context change hypothesis and is consistent with the idea that the interleaved retrieval events served to alter the context, leading each list to be contextually more distinct from the others.

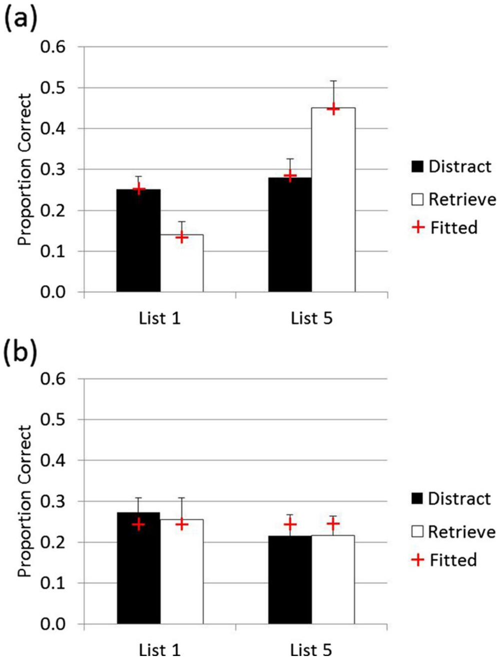

Fig. 1 Proportion recalled as a function of list (1 or 5) and task (distract or retrieve) for Experiment 1 (**a**) and Experiment 3 (**b**). The error bars represent the standard error of the mean. The model predictions (see Appendix 1) are indicated by the red cross

Analysis of intrusions

The context change hypothesis suggests that the memory advantage for list 5 in the retrieval condition is owing to reduced PI. Figure 2 shows the number of interlist intrusions during the final (free recall) test of lists 1 and 5. While inferential analysis was underpowered due to the low occurrence of intrusions, a mixed effects model including the fixed factors of intervening task and list, along with random intercepts for participants, was fit to the data (including random intercepts for the list from which the intrusions came did not reliably improve the fit of the model, $\chi^2 = 1.554$, $p = .213$). The model revealed a marginal main effect of list, $t = 1.951$, $p = .052$, with more intrusions in list 5; the main effect of task was not reliable, $t = 1.603$, $p = .1099$. The two-way interaction between task and list was also not reliable, $t = 1.404$, $p = .161$. However, analysis of the simple effects of task in each list revealed increased interlist intrusions on list 5 in the distractor condition, as compared with the retrieval condition, $t = 2.116$, $p = .035$, but no simple effect of task for list 1, $t = .141$, $p = .888$. This result suggests that PI accrued in the distractor condition but did not in the retrieval condition. The effect of retrieval on list 5 performance was likely driven by the reduction in the buildup of PI (no retrieval event occurred between studying list 5 and the final test, so the similarity of the test context and list 5 context remained consistent across both the retrieval and distractor conditions). The retrieval-driven reduction in PI from lists 1–4 led to overall enhanced performance on list 5.

Notably, no effect of interlist intrusions is apparent in list 1, suggesting that retrieval did not significantly influence RI. According to the context change hypothesis, reduced performance on list 1 in the retrieval condition was driven by the difference in contexts between list 1 and the final test, rather than enhanced discriminability between list 1 and later

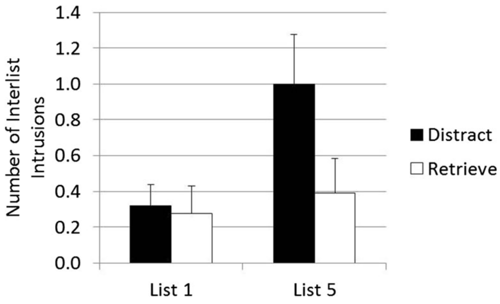

Fig. 2 Number of interlist intrusions as a function of list (1 or 5) and task (distract or retrieve) for Experiment 1. The error bars represent the standard error of the mean

potentially interfering lists. Because each interleaved retrieval event during the study session caused context to shift further than it normally would, the context at the time of the final test was further down the contextual stream (and thus more disparate than the list 1 context) in the retrieval condition than in the distractor condition. This greater mismatch of list 1 context and criterion test context led to overall *reduced* performance on list 1.

Experiment 2

The benefits of reducing PI by using a retrieval task hold some promise of aiding human learning in more educationally representative situations than simple list learning. Experiment 2 extends the methods used in Experiment 1 from word lists to text materials. The most important change that follows from this shift in materials concerns the nature of the test. Whereas the tests of recall implemented in the first experiment indicates the effects of retrieval on episodic memory, the test in Experiment 2 queries a more general understanding of complex materials by using multiple-choice and short-answer tests similar to those used in classroom testing.

Method

Participants

Twenty-eight undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment for course credit.

Design

Type of intervening task (unrelated semantic retrieval or nonmnemonic distractor control) and testing order of the studied texts were both manipulated between participants. The tested texts (1 and 4) were counterbalanced between participants, and the filler texts (2 and 3) were counterbalanced between participants independently. Memory performance was measured via short-answer and multiple-choice questions.

Materials

Text materials were drawn from a standardized test (ACT) prep book (Dulan 2010). The texts were related to animals (coyotes, porcupines, seals, and chronic wasting disease [CWD]) and averaged 608 words (*SD* = 55.62). The coyote and porcupine texts were in tested positions (text 1 or 4), while the seal and CWD texts were in filler positions (text 2 or 3).

Procedure

The procedure was similar to that of Experiment 1. The texts were presented one paragraph at a time. Reading time was self-paced, and participants could not go back to a previous paragraph after advancing to the next one. Participants studied a total of four texts. Each text was separated by an intervening task. Half of the participants were in the distractor condition (three sets of counting backward by 3s from a three-digit number for 30 s each), and the other half were in the retrieval condition (counting backward for 30 s, followed by 60 s of unrelated semantic retrieval). The semantic retrieval categories (listing types of sports, professions, or office supplies) were randomly ordered for each participant. After studying text 4, participants completed two more 30 s sets of the distractor counting task. The test was administered via pen and paper. For both text 1 and text 4, participants completed short-answer questions, followed by 10 multiple-choice questions (see Appendix 3 for the test materials used). As in the previous experiments, participants were tested on the alternative text following the first test. Participants were instructed not to go back and change answers after moving to the next question, and no time limit was given to complete the test.

Results and discussion

The mixed effects models reported here were fitted via Laplace estimation in a manner similar to that in Experiment 1. Only the results from the first test are reported here (see Appendix 2 for results and discussion of the second test). The analyses were separated on the basis of question type (multiple choice or short answer). Mean proportions correct are reported in Fig. 3a (multiple choice) and Fig. 3b (short answer).

Multiple choice

The best-fit model for the multiple-choice data included the fixed effects of text (1 or 4) and intervening task (retrieval or distractor). Including random intercepts for question, in addition to random intercepts for participant, significantly improved the fit of the model, χ^2 = 5.056, p = .0246. Having the random effect of question also allowed the model to take into account any variance introduced due to which text (e.g., coyote or porcupine) was being tested. While there was no reliable main effect of task, z = .063, p = .971, a main effect of text occurred in which overall performance on text 1 was higher than that on text 4, z = 2.507, p = .0122. Furthermore, a reliable two-way interaction revealed enhanced performance for text 4 following the retrieval task but reduced performance for text 1 following the retrieval task, z = 2.894, p = .00381. This pattern of results was supported by a marginal simple effect of task in text 1, z = 1.893, p = .0584, and a reliable

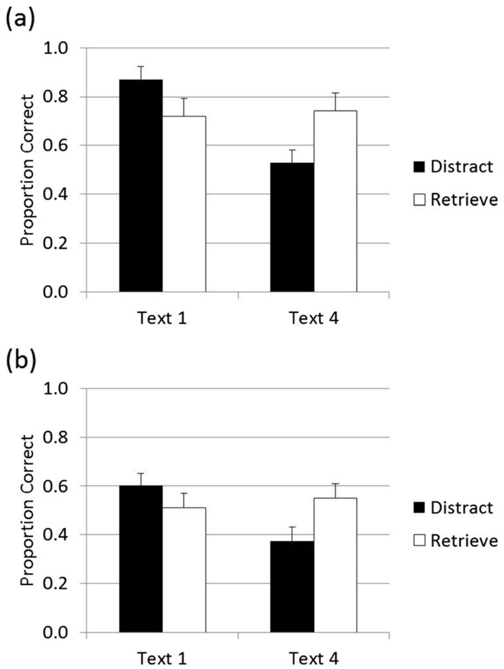

Fig. 3 Proportion correct as a function of list (1 or 5) and task (distract or retrieve) for Experiment 2 multiple-choice questions (**a**) and short-answer questions (**b**). The error bars represent the standard error of the mean

simple effect of task in text 4, $z = 2.236$, $p = .0254$. Notably, the data followed similar trends, regardless of text type (e.g., coyote vs. porcupine).

Short answer

Two independent raters scored the short answer questions (intraclass correlation, $r = .789$). The best-fit model included the fixed effects of intervening task (retrieval or distractor) and text (1 or 4). It also included random intercepts for participant, question, and rater, along with random slopes for question (the fit of the model was significantly improved by including random slopes instead of only random intercepts for question; $\chi^2 = 21.961$, $p = .009$). Because the usual Markov chain Monte Carlo method of determining reliability has not been implemented for this type of model, we instead calculated significance using "worst case scenario" estimates of the degrees of freedom ($df = 200$)[3] for the given t-values.

The results revealed no main effect of task, $t = 0.634$, $p = .527$, or text, $t = 1.337$, $p = .183$. However, a reliable interaction, $t = 2.035$, $p = .043$, indicated that participants in the retrieval task had lower performance on text 1 but higher performance on text 4 than did those in the distractor task. Analyses of the simple effects revealed that this trend was marginally reliable in text 4, $t = 1.881$, $p = .0614$, but not in text 1, $t = 0.982$, $p = .327$.[4] Although the short-answer test did not convincingly replicate the results from the multiple-choice analysis, it is worth noting that the patterns are the same on the two tests and appear to follow the predictions of the context-shift model and that all tests used very conservative degrees of freedom.

The multiple-choice results from Experiment 2 replicated the results of Experiment 1, extending the effect to materials more relevant for educational settings. Having an unrelated retrieval event between studying texts led to *reduced* performance on earlier material but *enhanced* performance on later material. Once again, these results suggest that interleaved retrieval events have the beneficial effect of segregating texts (and thus reducing intertext interference) but the harmful effect of reducing access to earlier material (due to a greater shift down the contextual stream). The effects here are especially impressive in light of the many cognitive skills a reader must coordinate in order to read, understand, and remember a complicated text.

Experiment 3

Experiment 3 was designed to examine the effects of delaying the final test after the study session. Introducing a long enough delay should lead to equivalent performance, regardless of the rate of contextual fluctuation during study. The context at test will be shifted sufficiently far away from the list contexts that the differences between those list contexts are relatively small (and therefore, the benefits and costs of list isolation due to greater context fluctuation are eliminated).

Method

Participants

One hundred seventeen undergraduate students at the University of Illinois at Urbana-Champaign participated for course credit. Data from 13 participants were dropped for not properly following the instructions (e.g., recalling items from previous lists when they should be listing items from the given semantic category).

[3] Since our data fell on the borderline of whether the t-distribution approximated the normal distribution, we instead calculated a conservative, "worst case scenario" estimate of the degrees of freedom. We took the number of observations and subtracted from it the number of fixed effects, the number of participants, the number of raters, and the number of questions times the number of fixed effects.

[4] On the basis of sample size and the results of the previous experiments, we examined effect sizes for the simple effect of condition in both text 1 ($\eta_p^2 = .092$) and text 4 ($\eta_p^2 = .251$). These effect size estimates are particularly conservative, since they do not factor in the influence of the random effects (e.g., rater, question, and participant).

Design, materials, and procedure

Experiment 3 was identical to that of Experiment 1, with the following exceptions. Presentation time for each word was increased from 4 s to 6 s. A 15-min delay was introduced between the study session (after the 60 s of counting following the final list) and the final tests. During the delay, participants performed a spatial-matching task. In the task, tokens of different colors were arranged in a grid, and a participant's job was to swap adjacent tokens in order to make chains of three or more tokens of the same color. This task was chosen because of its general contrast with the other tasks in the experiment and low reliance on verbal codes.

Results and discussion

Once again, the mixed effects models reported here were fitted via Laplace estimation in a similar fashion as in Experiments 1 and 2. The results are shown in Fig. 1b. Again, only the results of the first test are reported here (see Appendix 2 for results and discussion of the second test).

The best-fit model included the fixed effects of list (1 or 5) and intervening task (retrieval or distractor), along with random intercepts for participants. Adding random intercepts for item did not reliably improve the fit of the model, $\chi^2 = 1.845$, $p = .174$. While there was a marginal main effect of list with higher performance in list 1 than in list 5, $z = 1.786$, $p = .074$, hinting at a recency-to-primacy shift across Experiments 1 and 2, there was neither a reliable main effect of task, $z = 0.242$, $p = .809$, nor a reliable interaction between list and task, $z = 0.129$, $p = .897$.

The results from Experiment 3 indicate important boundary conditions on the results evident in Experiment 1. There was no effect of interleaved retrieval on list 1 and list 5 performance. This result was predicted a priori from a context change perspective. According to the context change hypothesis, the increased retention interval between the study session and the final test caused the context at test to shift sufficiently far away from the list contexts that the increased differentiation between list contexts (via interleaved retrieval) was now quite small, *relative* to the difference between those contexts and the criterion test context. Thus, list isolation no longer benefited performance. In addition, list 5 no longer received the benefit of sharing a similar context with the final test. Interleaved retrieval did not influence list 1 performance, because the difference between list 1 and final test contexts was not dramatically different across the retrieval and distractor conditions (relative to the new, very different criterion test context). Taken together, these results indicate that the effects of retrieval depend on the relative placement within the contextual stream of the criterion test context. However, other materials with different forgetting rates and different levels of relatedness (such as that used in Experiment 2), along with the extent of context change implemented (such as by varying duration), might show varying rates of extinction (e.g., in similar fashion to the endurance of recency effects; Bjork and Whitten 1974; Cepeda, Vul, Rohrer, Wixted and Pashler 2008). Experiment 3 takes the first step in highlighting important boundary conditions on the effect of semantic retrieval on prior and future learning. Further research on the type of material studied and the type and length of intervening and delay tasks, along with other possible explanations for the absence of an effect after a delay, is needed to further map out the influence of interleaved retrieval.

General discussion

The present study utilized a multilist learning paradigm to explore the beneficial and harmful effects of unrelated semantic retrieval on memory for both earlier and later items. Interleaved retrieval events led to overall *reduced* performance on items learned prior to retrieval but *enhanced* performance on items learned after retrieval. This effect held for both word lists and text materials (under multiple-choice testing) when the criterion test immediately followed the study session; however, the effect disappeared for text materials when the criterion test was delayed.

These results follow directly from the context change hypothesis of retrieval. The more similar the study context and the test context, the more likely one will be to remember items from the study session. Likewise, the more similar two study contexts are, the more likely they will be to interfere with each other. The interleaved retrieval events served to alter participants' internal context, causing it to shift more rapidly than it normally would. This led to the potentially beneficial effect of each list context in the study session being more differentiated from the others (and thus less likely to interfere with each other). However, the interleaved retrieval events (and subsequent context shifts) also led to the potentially harmful effect of the criterion test context being shifted much further away from the study contexts than would have normally occurred with no interleaved retrieval events.

The dual nature of retrieval events—beneficially segregating the study contexts but decreasing the match between study context and criterion test context—led to improved access for later learning but impaired access for earlier learning. Because the last study session (e.g., list 5) was not affected by an additional shift in context (via a retrieval event) between it and the criterion test, the driving factor for enhanced performance on the final list was likely the reduction in PI. This claim is supported by the analysis of intrusions in Experiment 1 and by a similar analysis by Szpunar et al. (2008), which revealed temporally graded and ample intrusions in the distractor condition but not many intrusions in the retrieval condition (thus, retrieval appears to reduce PI but not RI; see Fig. 2 in the present article and Fig. 4 in Szpunar et al. 2008). However, while a reduction in PI certainly appears to be one basis for the enhanced performance for later material, it might not account for the entire story (e.g., strategy shifts as in Wissman, Rawson and Pyc 2011, and Delaney and Knowles 2005; length of intervening task as in Unsworth, Spillers and Brewer 2012; enhanced encoding as in Pastötter et al. 2011). Other mechanisms for this effect of reduced intrusions and enhanced performance on later items that would be applicable to semantic as well as episodic retrieval have yet to be explored. Interestingly, RI did not appear to play a significant role in either condition: Interleaved retrieval led to overall reduced performance for earlier items (e.g., list 1), with no obvious shift in the pattern of intrusions.

One possible concern about these effects might be that the semantic retrieval task has a different level of difficulty than the distractor counting task. However, the level of difficulty does not appear to be a driving force here. Pastötter et al. (2011) found a similar magnitude of effect on future learning for three different types of retrieval tasks (where difficulty was not measured) and an equivalent (lower) magnitude of effect on future learning for a distractor counting task and restudy (also where difficulty was not measured)[5]. Furthermore, Sahakyan and Hendricks (2012) manipulated the difficulty of the retrieval task in a list-before-last paradigm and found no influence of degree of difficulty on the effect of retrieval on prior learning.

Introducing a delay between the study session and criterion test allowed the test context to shift further away from that of the study session. The context shifts (via interleaved retrieval) between the lists in the study session were small, *relative* to the change induced by the delay, since both effects disappeared in this condition. The results from the delayed test indicate an important boundary condition for the influence of interleaved retrieval within a study session. When the context at criterion test is sufficiently different from that of the study contexts, the retrieval-driven differences in context are relatively small. The present study indicates that a significant shift in context at criterion test can occur in as little as 15 min for word list materials. Events (other than time delays) that significantly change context are also likely to eliminate the effects of retrieval, although the artificial nature of the task in Experiments 1 and 3 does not provide much guidance on what length of delay or degree of intervening retrieval would be influential in real-world learning. The latter point is especially important when considering the significance of the findings of Experiment 2 (which used richer, more complex material and would thus likely need a greater delay to eliminate the effects of retrieval). It does appear that the context at criterion test must bear a certain degree of resemblance to that of the study session in order for interleaved retrieval to exert its effects. Further exploration of the boundaries of the influence of retrieval is still needed. It is possible, for example, that reinstating aspects of the study session through external contextual cues can enhance the effects of retrieval on memory, even at a more substantial delay. Additionally, further studies on the influence of delay before the criterion test should rule out alternative explanations to the context change hypothesis (such as other memory factors, like consolidation, playing a more prevalent role than context at long delays).

While further research is still needed to determine the boundaries of the effects of retrieval during study, many potential applications exist. Most prominent is the application to educational settings. While introducing context change (via unrelated semantic retrieval or another method) during study enhances later learning, it also impairs retention for material learned earlier. Moreover, this effect extends from rote episodic list-learning to more general understanding and memory for complex text materials. In contexts in which PI is a major source of difficulty—for example, with older adults (Lustig, May and Hasher 2001) or with materials that lend themselves to such difficulty (e.g., conceptually similar materials or those learned in similar contexts)—interleaved retrieval and the attendant trade-offs between early- and late-list learning might yield overall positive effects. While the boundaries of these potential applications (e.g., influence of delay or context reinstatement) need to be further explored, awareness of how simple retrieval tasks might induce context change and how that, in turn, affects learning is the first step to enhancing classroom learning.

The present set of experiments examined the negative and positive consequences of interleaved unrelated semantic retrieval within a study session. When the criterion test immediately followed the study session (and thus held some resemblance to the study contexts), retrieval led to *enhanced* performance for later material via a reduction in the buildup of PI but overall *reduced* retention for earlier material. While this effect held for both word lists and more complex text materials, it disappeared when the final criterion test was delayed for word list materials. Further research is still necessary to deepen our understanding of the potentially beneficial and harmful effects of retrieval and their applications.

[5] One other concern the reader might have, at first glance, is that the retrieval task (generating words) is more similar to the materials used in Experiment 1 (lists of unrelated words) than is the distractor task (counting backward). If this effect were driven by the interleaved task being word based or number based, then Pastötter et al. (2011) should have found enhanced performance for the tasks involving words (testing, restudying, generation) but not numbers (counting, *n*-back); however, it was the *retrieval* nature of the task (testing, generation, *n*-back), not the material type, that mattered. Also, the text materials used in Experiment 2 are relatively less similar to the generation task, yet the effect still remained when multiple-choice questions were asked (despite a significantly smaller sample size).

Acknowledgments This research was funded in part by Grant R01 AG026263 from the National Institutes of Health and was conducted for the first author's master's thesis at the University of Illinois at Urbana-Champaign. We thank the members of the Human Memory and Cognition Laboratory at the University of Illinois at Urbana-Champaign for their helpful insights, comments, and criticism.

Appendix 1

Many excellent models exist that could explain components of our experiments. For example, Jang and Huber (2008) developed a model that took into consideration context similarity, context retrieval, and censorship of the intervening list, which worked within the framework of their list-before-last experiments. Importantly, this was implemented to consider the influence of interference from competing lists and list isolation. Another relevant model is that of Lehman and Malmberg (2009), which explains both unintentional and intentional forgetting via contextual interference.

While these are models of note and may be of interest to the reader, our goal here was to create a simple tool that focused just on the underlying core concept of our theory—that speeded context fluctuation (via interleaved retrieval) will lead to enhanced performance on later items but reduced performance on previous items.

In order to evaluate these intuitions, a simple mathematical model was developed. The core assumptions of the model are taken from Estes' (1955) stimulus sampling theory (SST). SST posits that at any given time, a select number of contextual elements is available. The elements are drawn from a sample of the population of all possible elements. The set of elements not available are distinguished from those currently available. Importantly, Estes proposed that the elements randomly fluctuate between the available and unavailable sets. The rate of fluctuation is represented by *a*, and the proportion of available elements for encoding is represented by *J*. As the time interval increases, so does the amount of fluctuation. Estes originally applied SST to animal learning (and conditioned responses), but it is also applicable to human memory and learning (e.g., Bower 1967). The probability of an element appearing in two contexts separated by a time interval *I* is

$$p(e \text{ in } T_1 \text{ and } T_2) = J + (1–J)a^I,$$

where T_1 and T_2 represent the first and second contexts and *I* is the interval between them (Benjamin and Tullis 2010).

In the present set of experiments, access to multiple lists must be accounted for, and a mechanism for competition between those lists must be implemented. To simplify the situation, we will consider only the first and the last lists in a multilist paradigm (although the logic can be easily generalized to any number of lists). In addition, we will implement competition by assuming that the probability of accessing a given memory trace is a ratio of the number of elements uniquely encoded with that stimulus and present at test to the number of elements encoded in both the sought-after and the interfering trace and present at test (factoring in how unique the match is better reflects the process of retrieval; Nairne 2002; Poirier et al. 2012). The probability that an element is in both the target list (L_T) and the criterion test (*T*), taking into account interference from the other list (L_O) is then given by

$$\frac{p(e\ \ in\ \ T\ and\ \ L_T)}{p(e\ \ in\ \ T\ and\ L_T\ \ or\ L_o)} \tag{1}$$

$$= \frac{J+(1–J)a^{I_T}}{[J+(1–J)a^{I_T}]+[1–[J+(1–J)a^{I_T}]][J+(1–J)a^{I_O}]}$$

where I_T represents the interval between the target list and the criterion test and I_O represents the interval between the interfering list and the criterion test.

The hypothesis being tested within this model is that retrieval events serve to enhance context fluctuation. In order to evaluate the potential of the model to yield the predicted effects using only this variable, *a* was varied ($a_\text{d} = .1$, $a_\text{r} = .7$, where a_d and a_r represent the rates for distractor control and retrieval conditions, respectively) while holding the other variables constant (J = .3; I_T = 1 and I_O = 5 when the later list is the target; I_T = 5 and I_O = 1 when the earlier list is the target). The

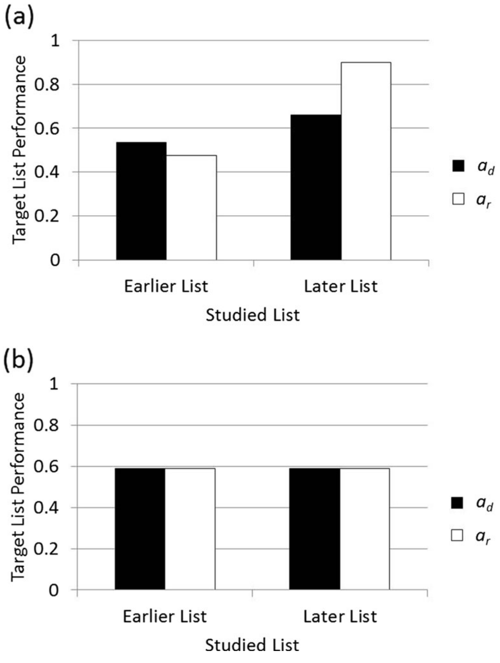

Fig. 4 Simulated proportion recalled on target list as a function of list and *a* (parameter for the amount of fluctuation; a_d refers to the distractor condition and a_r to the retrieval condition) for Experiment 1 (**a**) and Experiment 3 (**b**)

results of the simulation are shown in Fig. 4a. As predicted by the context change hypothesis, greater fluctuation (e.g., retrieval) led to enhanced performance on the later list but impaired performance on the earlier list, as compared with the condition with less fluctuation (e.g., control). We also simulated the influence of introducing a long delay before the final test. We implemented this change in the model to include this delay ($d = 20$) by adding it to the parameters of I_T and I_O. The results of the simulation are shown in Fig. 4b. As was expected, the delay alleviated the effects of retrieval.

Model fit of Experiment 1

Equation 1 was fit to the data from Experiment 1. As in the simulation, the parameter J was set to 0.3. When the target list was list 1, the parameters I_T and I_O were set to 5 and 1, respectively; when the target list was list 5, the parameters I_T and I_O were set to 1 and 5, respectively. The model was augmented to include an additive scaling constant (s) that was added to performance in every condition. a was allowed to vary freely on the basis of intervening task, and s varied freely, yielding three free parameters for four data points. The model was fit via least squares estimation to mean performance from Experiment 1 (a_d = .022, a_r = .293). The red crosses in Fig. 1a show the fit of the model to the data from Experiment 1. The model replicated the data nearly perfectly, yielding identical fits to four significant digits (SS = .0000206; RMSD = .00227). To address concerns about model validity, the model was also estimated from a random half of the data and fitted against the other half. Least squares estimation in that case also revealed an excellent fit (SS = .00238; RMSD = .0244).

Model fit of Experiment 3

The a_d and a_r values from the fitted model from Experiment 1 were used to fit a model to the data from Experiment 3. A delay (d) was introduced to the intervals between each list. This delay and the scaling factor were both allowed to vary freely, yielding two free parameters. The model was fit via least squares estimation on mean performance from Experiment 3 (d = 15.054). The model fit is shown by the red crosses in Fig. 2b. The model replicated the data quite well, with identical fits to two significant digits (SS = .0017; RMSD = .0206). Once again, to address concerns of validity, the model was also estimated from a random half of the data and fitted against the other half. Least squares estimation once again revealed an excellent fit (SS = .0126; RMSD = .0562).

Appendix 2

The first test of the critical material (e.g., list 1 or list 5) influenced performance on the second test. The results of the second test across experiments are reported here. The same methods were used as in the main analysis of the first test. Except where noted, the same models created the best fit. See Fig. 5 for mean proportion correct for all three experiments.

Experiment 1

(Note that the best fit model included random intercepts for item in addition to participant.) There was no reliable main effect of task, z = 0.914, p = .360, or list, z = 0.693, p = .488. The two-way interaction between task and list also was not reliable, z = 1.539, p = .124.

Experiment 2

Analysis of the multiple-choice responses revealed a marginal main effect of task with overall higher performance in the distractor condition, z = 1.801, p = .0717, but no main effect of text, z = 0.630, p = .529. The two-way interaction between task and text was not reliable, z = 1.547, p = .122. Analysis of the short-answer responses revealed no main effects of task, t = 0.230, p = .818, or text, t = 0.919, p = .359, and no two-way interaction between task and text, t =0.960, p = .338.

Experiment 3

The analysis revealed a main effect of list with overall higher performance in list 1, z = 5.200, p < .001, but no main effect of task, z = 0.191, p = .848. A significant two-way interaction between task and list, z = 3.612, p < .001, revealed enhanced performance on list 1 but reduced performance in list 5 for participants in the retrieval task, as compared with the distractor task. One possibility is that the reinstatement induced by the first test of the material shifted the current context back to the study context (therefore alleviating the effect of the delay).

Appendix 3

Short-answer questions (porcupine text):

1. Describe how a porcupine uses its quills for defense AND the consequences for its victim.
2. Describe the negative AND positive opinions held about porcupines in different regions of the world.
3. Describe a porcupine's typical habitat and daily routine.
4. Describe a porcupine's defense system AND how it might be overcome by a predator.
5. Describe the porcupine's cycle of reproduction AND characteristics of the young.

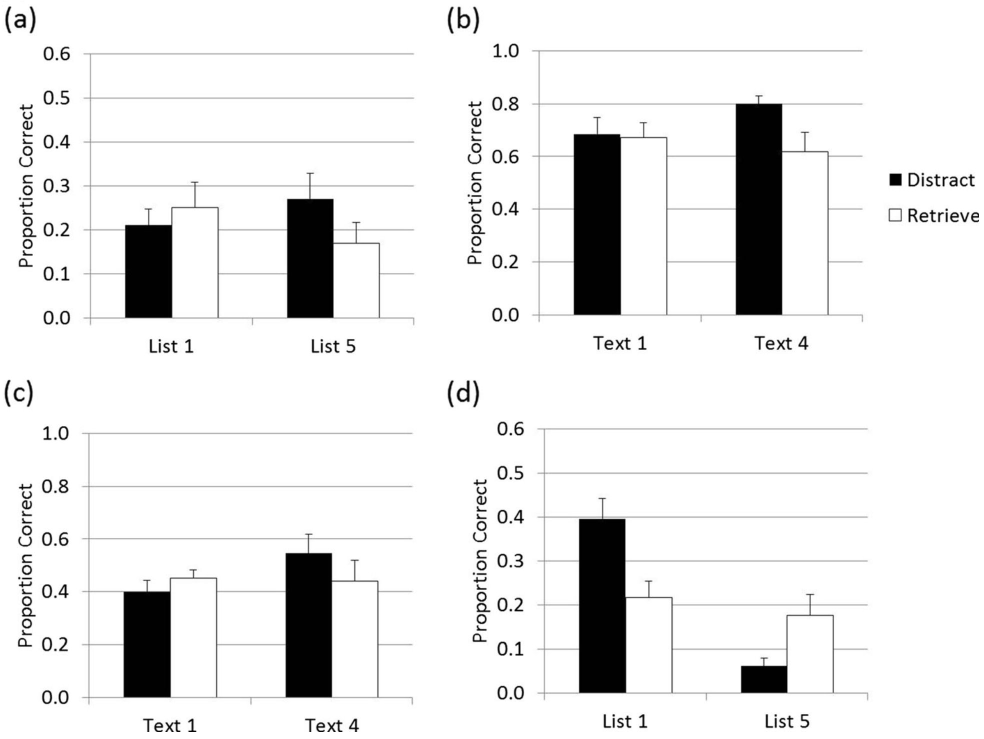

Fig. 5 Proportion correct as a function of study position (list 1 or list 5 for Experiments 1 and 2; text 1 or text 4 for Experiment 2) and task (distract or retrieve) for the second criterion test in Experiment 1 (**a**); Experiment 2, multiple-choice questions (**b**); Experiment 2, short-answer questions (**c**); and Experiment 3 (**d**). The error bars represent the standard error of the mean

Multiple-choice questions (porcupine text):

1. How many quills does a typical porcupine have?
   a. 10,000
   b. 20,000
   c. 30,000
   d. 40,000

2. How large is a typical adult porcupine (in length)?
   a. 1–2 ft
   b. 2–2 ½ ft
   c. 1 ½–3 ft
   d. 2–3 ½ ft

3. How long is the porcupine's gestation period?
   a. 4 months
   b. 5 months
   c. 6 months
   d. 7 months

4. Which of the following were NOT identified as successful predators of porcupine?
   a. Dogs
   b. Bobcats
   c. Cougars
   d. Coyotes

5. What regions do porcupines generally inhabit?
   a. Northern
   b. Eastern
   c. Southern
   d. Western

6. How do porcupines USUALLY sleep?
   a. On their backs
   b. In a tree
   c. Burrowed underground
   d. Under dense foliage on the ground

7. What use for porcupine quills was NOT listed in the passage?
   a. Jewelry
   b. Hair accessory
   c. Clothing
   d. Shoes

8. How many offspring does a female typically give birth to in a year?
   a. 1
   b. 2
   c. 3
   d. 4

9. In terms of size, where does the porcupine rank in the rodent family?
   a. 1st
   b. 2nd
   c. 3rd
   d. 4th

10. What do porcupines like to eat?
    a. Beaver
    b. Small rodents
    c. Bark
    d. Needles

Short-answer questions (coyote text):

1. Describe the characteristics commonly attributed to coyotes in stories AND where they originated.[6]
2. Compare AND contrast characteristics of the coyote and collie.
3. Describe the coyote's eating habits in terms of both sustenance and environment. Give examples.
4. Describe the coyote's hunting patterns and how and when they might vary.
5. Describe and give examples of the environments coyotes might inhabit.
6. Describe and give examples of coyote's physical abilities.

Multiple-choice questions (coyote text):

1. What types of prey will coyotes hunt when working in a team?
   a. Small pets
   b. Deer
   c. Sheep
   d. Young livestock

2. What specific animal was mentioned as a hunting partner?
   a. Beaver
   b. Badger
   c. Cougar
   d. Collie

3. How fast can coyotes run?
   a. 10 mph
   b. 20 mph
   c. 30 mph
   d. 40 mph

4. What's the HIGHEST a coyote can leap?
   a. 8 ft
   b. 11 ft
   c. 14 ft
   d. 17 ft

5. What feature was NOT mentioned as something coyotes are willing to overcome?
   a. Swimming
   b. Urban environments
   c. Cyclone fences
   d. Other large predators

6. What are the coyotes yips used for?
   a. Frighten prey
   b. Mating calls
   c. Warning signals
   d. General communication

7. Where did the coyote originate?
   a. Northern US
   b. Eastern US
   c. Southern US
   d. Western US

8. What was NOT mentioned as being done to keep the coyote population in check?
   a. Trap
   b. Shoot
   c. Increase natural predators
   d. Poison

9. Where are coyotes NOT found
   a. North America
   b. Central America
   c. South America
   d. Arctic

10. Which of the following were not mentioned as natural predators to the coyote?
    a. Wolves
    b. Mountain lions
    c. Bears
    d. Badgers

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Bates, D., Maechler, M., & Dai, B. (2011). Lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Retrieved

[6] This question was not used in the analyses due to multiple participants misunderstanding to what it referred

from http://lme4.r-forge.r-project.org (R package version 0.999375-39)

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247.

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

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6,* 173–189.

Bower, G. (1967). A multicomponent theory of the memory trace. *Psychology of Learning and Motivation, 1,* 229–325.

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102.

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

Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects in the free recall of unmixed lists. *Journal of Memory and Language, 52,* 120–130.

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

Dulan, S. W. (2010). *McGraw-Hill's ACT* (5th ed.). Dubuque: World Color.

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

Glanzer, M. & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 351–360.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446.

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

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

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting multiple lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 970–988.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130*(2), 199–207.

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

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

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

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

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

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

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

Poirier, M., Nairne, J. S., Morin, C., Zimmermann, F. G. S., Koutmeridou, K., & Fowler, J. (2012). Memory as discrimination: A challenge to the encoding-retrieval match principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 16–29.

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

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

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

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

Sahakyan, L., Delaney, P. F., & Goodmon, L. B. (2008). Oh, Honey, I already forgot that: Strategic control of directed forgetting in older and younger adults. *Psychology and Aging, 23*(3), 621–633.

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

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

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

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

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

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

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

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

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012). Evidence for noisy contextual search: Examining the dynamics of list-before-last recall. *Memory, 20*(1), 1–13.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin and Review, 18,* 1140–1147.