Memory
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pmem20

Testing improves true recall and protects against the build-up of proactive interference without increasing false recall

Ludmila D. Nunes [a b] & Yana Weinstein [b]

[a] Department of Psychology, University of Lisbon, Lisbon, Portugal

[b] Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

Available online: 31 Jan 2012

To cite this article: Ludmila D. Nunes & Yana Weinstein (2012): Testing improves true recall and protects against the build-up of proactive interference without increasing false recall, Memory, 20:2, 138-154

To link to this article: http://dx.doi.org/10.1080/09658211.2011.648198

MEMORY, 2012, 20 (2), 138–154

Psychology Press
Taylor & Francis Group

Testing improves true recall and protects against the build-up of proactive interference without increasing false recall

Ludmila D. Nunes[1,2] and Yana Weinstein[2]

[1]Department of Psychology, University of Lisbon, Lisbon, Portugal
[2]Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

Retrieval practice has been shown to protect against the negative effects of previously learned information on the learning of subsequent information, while increasing retention of new information. We report three experiments investigating the impact of retrieval practice on false recall in a multiple list paradigm. In three different experimental designs participants studied blocks of interrelated words that converged on non-presented associates. Participants were tested either after every study block or only after the fifth study block, and both groups received a cumulative test on all five study blocks. Overall the results from all three different experimental designs point to a benefit of testing in increasing long-term veridical recall on the cumulative test. More importantly, this improvement in veridical recall did not come at a cost: False recall on the cumulative test did not increase from retrieval practice.

Keywords: Testing; False memory; False recall; Proactive interference.

Testing has recently been proposed as an effective way to promote memory and learning (for a review see Roediger & Karpicke, 2006a). It has been widely demonstrated that after studying a set of materials (e.g., word lists), one is more likely to remember these materials on a later memory test following retrieval practice, i.e., taking multiple tests (e.g., Roediger & Karpicke, 2006b), even when the control condition involves re-exposure to the studied material (Karpicke & Roediger, 2008). Szpunar, McDermott, and Roediger (2008) demonstrated another benefit of testing during study—the prevention of the build-up of proactive interference (PI). The build-up of PI occurs when a memorisation task involves successive sets of materials that share the same category or modality (Keppel & Underwood, 1962; Postman & Keppel, 1977; Wickens, Born, & Allen, 1963). That is, learning the initial sets of materials interferes with learning later material, and therefore the retention of new information is inversely related to the number of prior learning sets (Underwood, 1957). To examine the impact of testing during study on the build-up of PI, Szpunar and colleagues (2008) had participants study five lists of words in anticipation of a cumulative test, with half of the participants tested after each list and the other tested only after the fifth list. For both interrelated and unrelated lists, participants who had been tested after studying each of lists 1–4 were better at

Address correspondence to: Ludmila D. Nunes, Faculdade de Psicologia, Alameda da Universidade, 1649-013 Lisbon, Portugal. E-mail: lsdnunes@gmail.com

Support for this research was provided by a James S. McDonnell Foundation 21st Century Science Initiative grant: a Bridging Brain, Mind and Behavior/Collaborative Award; and a fellowship from the Portuguese Foundation for Science and Technology to the first author. We would like to thank Kathleen McDermott, James Nairne, Jeff Karpicke, and the members of the Memory & Cognition Lab for their useful comments on this paper.

http://www.psypress.com/memory http://dx.doi.org/10.1080/09658211.2011.648198

learning the fifth list and produced fewer prior-list intrusions (i.e., words that had actually appeared in previous lists) than participants who had not been tested after each list. This effect was unrelated to re-exposure, as another group who restudied the lists achieved exactly the same accuracy as the control group who performed a distractor task after each list (Szpunar et al., 2008). This beneficial effect has since been replicated in another lab (Pastötter, Schicker, Niedernhuber, & Bäuml, 2011), and with a cued recall task (Weinstein, McDermott, & Szpunar, 2011).

FALSE MEMORY

In the present paper we apply Szpunar and colleagues' (2008) finding to associative memory errors that are produced when participants study lists that converge on semantic associates (Deese, 1959; Roediger & McDermott, 1995). The question addressed in this paper is whether the benefits of testing for learning multiple lists of associated words might come at a cost. Theoretically the main goal of this research is to gain information on the boundary conditions of the testing effect, using study materials that may be prone to negative effects of testing. Specifically we investigated the potential increase in associative memory errors (i.e., false recall) that may occur in tandem with increased veridical recall of certain types of information.

In the Deese (1959)/Roediger/McDermott (Roediger & McDermott, 1995; DRM) paradigm participants study lists composed of words related to a critical non-presented word (for example, participants could study the words *bed*, *rest*, *awake*, *tired*, etc., all semantically associated to the critical non-presented word *sleep*). This critical word is then spontaneously produced on recall tests, sometimes as often as words that had actually been studied (Roediger & McDermott, 1995). One explanation for this false memory effect is provided by the activation-monitoring framework. According to this account, activation automatically spreads from the presented associates to the non-presented critical word (Collins & Loftus, 1975; Roediger & Gallo, 2004). False recall of the critical word thus arises from this automatic spreading of activation in association with a failure of reality monitoring (Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981), which is needed to distinguish between internally and externally generated information.

TESTING AND FALSE MEMORY: REPEATED LISTS

Despite all the positive press surrounding testing, the effects of testing on false recall are somewhat in doubt.[1] McDermott (1996, Exp. 2) first studied the effects of multiple study/test trials on the level of false recall. The experiment consisted of five presentations of the same study list (composed of three DRM lists), with each presentation followed by a free recall test. Participants showed a significant decrease in the overall level of false recall of the critical words across trials. One day later participants performed a final free recall test. False recall on this final delayed test was higher than that on the fifth test taken on the previous day, suggesting that the benefit of testing in reducing false recall that was observed on the immediate tests diminished over time.

In a different design McDermott (2006) presented participants with three sets of six DRM lists, with each set tested either zero, one, or three times. On a final free recall test the number of previous tests was positively related to the likelihood of recalling studied words (a positive effect of testing), but also positively related to the likelihood to recalling non-presented critical words (an associated cost of testing). However, false recall was only increased by testing when comparing participants tested zero times with participants tested one or three times, and participants tested one or three times did not appear to differ significantly in the level of false recall on the final test (although this comparison was not reported). It is thus somewhat unclear from those experiments whether repeated testing increases false recall, beyond the effect of taking a single initial test.

The idea that false recall should increase in tandem with veridical recall is suggested by the spreading activation theory (Collins & Loftus, 1975; Roediger & McDermott, 1995), according

[1] Roediger and McDermott's (1995) original paper introducing the paradigm showed that taking a free recall test before a recognition test on DRM lists increased the level of correct and false recognition, although further investigation somewhat disputed this (see Roediger, McDermott, & Robinson, 1998, for a review). For the purposes of this paper we will focus on the effect of prior recall tests on a later free recall test, and not on recognition.

to which activation spreads from studied words to non-presented associated words. Thus, if veridical memory increases due to repeated reactivation of the study list during testing, the level of activation of the critical non-presented words should also increase, leading to a higher level of false recall. According to this view more tests lead to more activation of the presented words and, consequently, more spreading activation to the critical non-presented words. The two studies reviewed above suggest that while testing certainly increases veridical memory, this benefit may come with the associated cost of increasing false recall in some situations. However, McDermott (2006) pointed out that the benefits of testing in this paradigm outweighed the costs in that the increase in veridical recall was greater than the increase in false recall.

TESTING AND FALSE MEMORY: NON-REPEATED LISTS

Our set of experiments differs from the research described above in that we are examining the effects of testing one list (or multiple lists) on the learning of another. One other article we are aware of that has examined such a situation with respect to false recall is Chan and Langley (2010), who examined the effects of retrieval practice on eyewitness memory. In their paradigm participants initially studied an eyewitness episode and were either tested or not on that information before being exposed to misinformation. The authors found that taking a test on the original information led to an enhanced susceptibility to misinformation. The authors suggested that this might have arisen due to the protective effects of testing against PI, which enhanced the learning of misinformation presented after initial testing of the original event. Thus testing may not only increase false recall of information related to tested lists (McDermott, 1996, 2006), but also promote false recall of information related to new lists encoded after initial testing.

With regard to our paradigm we already know from Szpunar et al.'s (2008) work that taking a test after each list produces a release from PI and increases veridical recall on subsequent lists as compared with a condition where no such prior test occurs. However, previous research into the effects of testing on false memory (Chan & Langley, 2010; McDermott, 1996, 2006) suggests that this improvement in veridical recall may come at a cost if lists that produce strong associations to non-presented words are studied. Testing may thus increase not only veridical recall, but also false recall via increased backward associative strength (BAS), known for being the main predictor of false memories (Roediger, Watson, McDermott, & Gallo, 2001).

THE PRESENT EXPERIMENTS

In order to examine the effects of testing on the build-up of false memories across lists, the materials in the first reported experiment were DRM lists split across multiple study lists, which we will call *study blocks* throughout the article to distinguish them from the DRM *lists* from which the items were taken. A set of five study blocks was developed, each of them composed of a different subset of the same six DRM lists. Participants took a recall test either after each study block, or only after the fifth study block; all participants then took a cumulative recall test (see Figure 1 for a schematic of the procedure and materials). The level of false recall of the six critical words on both the fifth study block test and the cumulative test was compared for the tested and untested groups.

In a second experiment, instead of DRM lists split across blocks, we presented four study blocks, each of which consisted of 12 words from the same DRM list, followed by a fifth study block composed of three additional associates from each of the four presented DRM lists (see Figure 2). We used this manipulation in an attempt to replicate Experiment 1 in a design that increased the likelihood of false recall and more closely resembled Roediger and McDermott's (1995) procedure of presenting pure DRM lists. According to Robinson and Roediger (1997), when more associates are studied, the likelihood of false recall increases because activation of the critical word is increased via automatic spreading activation from the studied associates. For this reason, presenting more associates sequentially before each test should have increased activation of the critical word; our experimental design permitted us to look at whether taking a test increased or decreased that activation.

In a third experiment we again presented each DRM list in a blocked fashion (as in

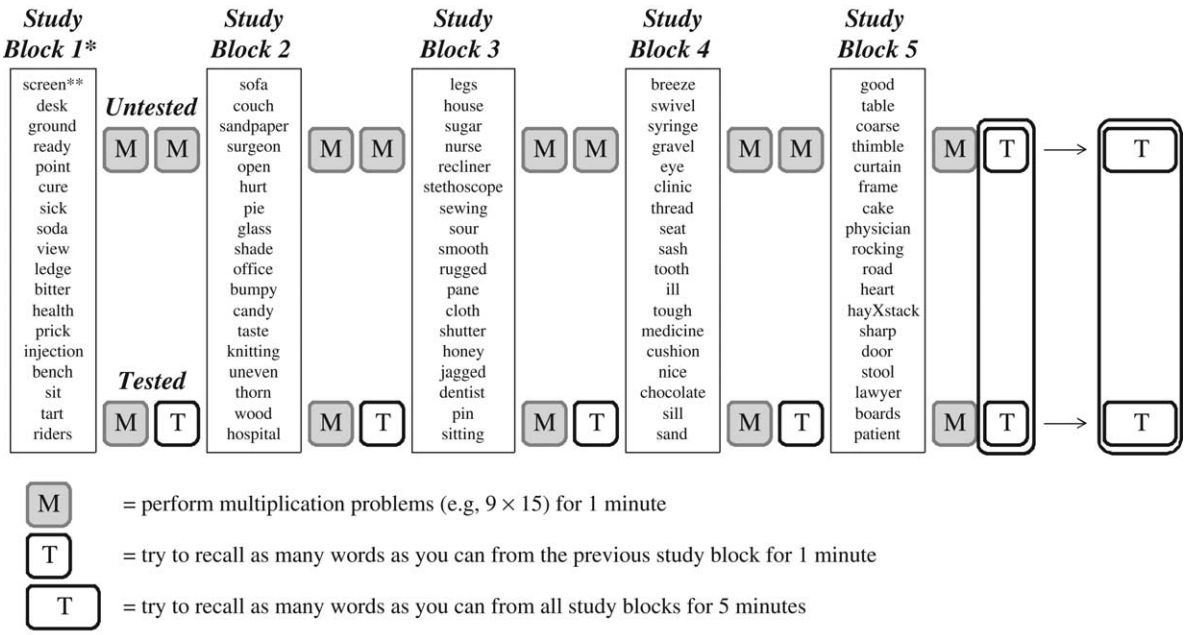

Figure 1. Schematic of the procedure and materials used in Experiment 1.

Experiment 2), but now these DRM lists were divided into three-word study blocks such that participants in the tested group received tests after each sub-list of three words and all participants were tested on the fifth and last sub-list of each DRM list (see Figure 3). This last experiment allowed us to examine the effect of testing on the build-up of false memories in a design that most closely followed the standard DRM paradigm.

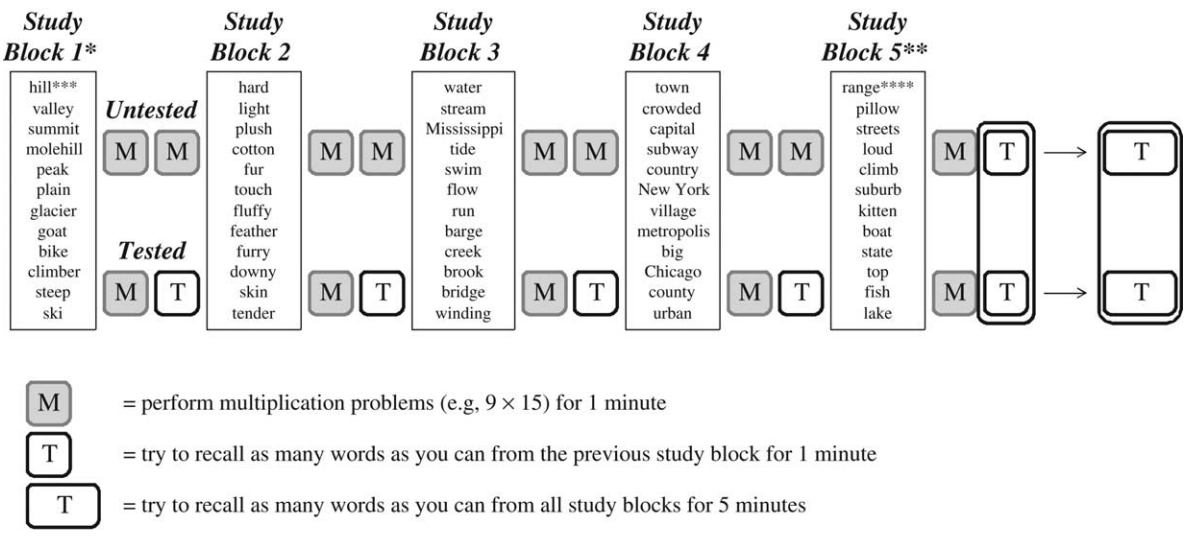

Figure 2. Schematic of the procedure and materials used in Experiment 2.

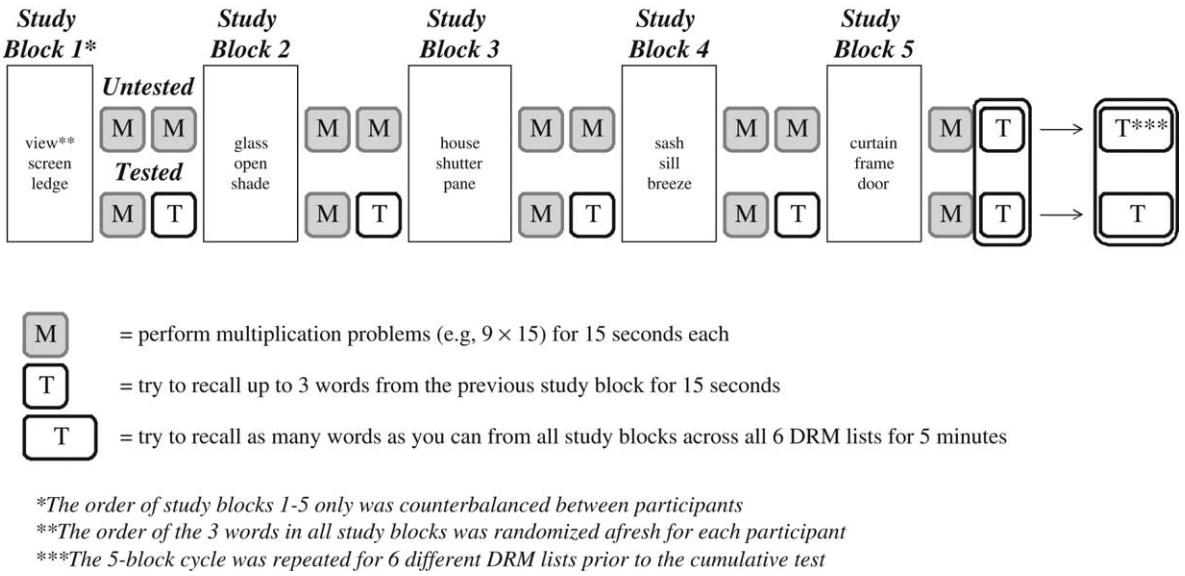

Figure 3. Schematic of the procedure and materials used in Experiment 3. Note that the schematic represents only one of the six DRM list cycles.

EXPERIMENT 1

Experiment 1 was designed to investigate the possible effects of testing on the build-up of false memories across multiple study blocks of categorised words, where each of the six categories represented a DRM list that was split across the five study blocks. The design of this experiment is identical to the one employed by Szpunar and colleagues (2008) to study the effects of testing on the build-up of PI in all but materials. The number of words studied in each list, and the number of participants in each experimental condition, are also based on Szpunar and colleagues' method (2008). Szpunar and colleagues' procedure involved five study blocks and a final cumulative test for all participants. Half of their participants were tested after every study block while the other half were tested only after the fifth study block. Performance on the fifth study block test and on the cumulative test was examined, comparing participants in the tested and untested groups. The only difference between our procedure and theirs is that we used lists of associated words that converged on critical words (i.e., DRM lists) instead of semantic categories. Szpunar and colleagues found that testing earlier study blocks helped to protect against prior block intrusions on the fifth study block test. Szpunar and colleagues also found that on the cumulative test, participants that had been tested after every study block produced overall more words from all five study blocks. Of course, PI could no longer be measured on the cumulative test. In our design interference in the form of false recall of critical words was measured both on the fifth study block test and on the cumulative test.

Method

Participants. A total of 62 Washington University undergraduates participated in the experiment for course credit or payment. Half of the participants were assigned to the tested condition, and the other half to the untested condition.

Materials. Five study blocks of 18 words were constructed from the DRM lists whose critical words were *chair*, *doctor*, *needle*, *rough*, *sweet*, and *window*. These six DRM lists were chosen because they were the most efficient in eliciting false recall, according to Stadler, Roediger, and McDermott's (1999) norms ($M_{\text{false recall}} = 56.3$). Each DRM list was randomised and divided into five groups of three words. Triads from each of the six DRM lists were then selected to form one study block, and this was repeated to make five study blocks in total. Thus each study block was made up of three randomly selected words from each of six DRM lists, so that presentation of the words from each DRM list was not ordered by BAS or any other criterion (see Figure 1 for the full set of stimuli as they were presented to participants).

Design. The design was between participants, so participants were either tested after study blocks one through five (tested group), or only after study block five (untested group; see Figure 1 for a schematic of the procedure in each condition). Performance between these two conditions was compared on two free recall tests: the test after the fifth study block, and the cumulative test after all five study blocks (the circled tests on Figure 1). The cumulative test was either delayed by 25 minutes or immediate, also between participants.[2] For the fifth study block test the levels of correct recall, false recall, and intrusions were measured. For the cumulative test the levels of correct recall and false recall were measured.

Procedure. Following the procedure used by Szpunar and colleagues (2008), participants were first informed that they would study five short blocks of words for a later test, and perform maths problems between the study blocks. Participants were additionally told that they might or might not receive a test after each individual study block, and that this would be determined randomly by the program. Participants were reminded to pay close attention because at the end of the experiment they would be tested on all five study blocks. In reality the testing schedule was determined by the program such that half of the participants received a free recall test after each study block and the other half only received a test after the fifth study block. All participants received a cumulative free recall test.

Words were presented one by one in the centre of a computer monitor at the rate of 2000 ms per word (500 ms interstimulus interval, 2500 ms stimulus onset asynchrony). Study blocks were counterbalanced using a Latin square design, resulting in five different study block orders for each of the testing groups. Words within each study block were randomised afresh for each participant. However, the assignment of words to blocks was constant between participants. After presentation of each study block, all participants solved maths problems for 1 minute. Tested participants then had 1 minute to complete a free recall test, in which they were instructed to type as many words as possible from the block they had just studied; meanwhile, untested participants performed another minute of maths problems. Following the fifth study block and 1 minute of maths problems, participants in both conditions were instructed to recall as many words as possible from that block. Participants were then given 5 minutes to recall, in any order, all the words from all of the study blocks (note that this test occurred immediately for some participants and after a 25-minute distractor task for others; see footnote 2).

[2] We manipulated the delay between the fifth study block test and the cumulative test to determine whether the effects of testing on false recall differed as a function of retention interval. This manipulation was driven by previous researching showing that initial testing can produce different results when the effects are assessed at different retention intervals (Roediger & Karpicke, 2006a). However, since this manipulation did not mediate the effects of testing on veridical or false recall, for the sake of brevity we do not discuss it further.

Results

Analyses of the results showed no differences between the delayed and the immediate test conditions on any measures, so this factor is not considered further. Two untested participants were excluded from the analyses because they produced more intrusions than studied words on the cumulative test. One tested participant was excluded from the analyses because they produced only one word on the fifth block test. We report correct recall and intrusion rates separated into prior block intrusions (words presented in previously studied blocks) and critical intrusions (critical words from each DRM list used). Following Szpunar et al. (2008), extra-list intrusions (any words that were not presented on prior blocks nor considered critical words) are not reported. All data are presented in terms of proportions. The denominators from which proportions were calculated for each item type are as follows. For the fifth study block test: correct recall out of 18 (the number of words presented in the fifth study block); critical words out of 6 (the number of DRM lists used in the experiment); and prior-list intrusions out of 72 (the number of words presented in study blocks one through four). For the cumulative test: correct recall out of 90 (the total number of words presented across the five study blocks); and critical words out of 6 (the number of DRM lists used in the experiment). The Alpha level was set at .05. Cohen's *d* indicates effect size for *t*-tests.

Fifth study block test. The mean proportion of correctly recalled words and critical words produced on the fifth study block test in each

condition are presented in Table 1. Participants who had not been tested on previous study blocks recalled fewer words from study block five than participants who had been tested after every study block (see top row of Table 1); $t(57) = 2.94$, $d = 0.77$. On the other hand, the level of false recall (mean number of critical words recalled out of six possible words) was higher for participants who had not been tested on the previous blocks (see second row of Table 1); $t(57) = 3.74$, $d = 0.96$. That is, 12 of the 29 untested participants produced either one (8 participants) or two (4 participants) of the six possible critical words on the fifth study block test. Meanwhile, only 1 of the 30 tested participants produced a critical word on the fifth study block test. We also calculated the mean proportion of critical words produced across the five study block tests for tested participants ($M = .07$). This did not differ significantly from the mean proportion of critical words produced on the fifth study block test in the group that had not received prior tests ($M = .09$).

We also compared the proportion of prior block intrusions produced on the fifth block test in each condition. Untested participants produced far more prior block intrusions than those who had been tested after every block ($M = .07$ vs .01); $t(57) = 4.97$, $d = 1.29$. These prior block intrusion data are consistent with those presented by Szpunar et al. (2008).

Cumulative test. The mean proportions of studied words and critical words produced on the cumulative test in each condition are also presented in Table 1. In line with Szpunar et al. (2008) we found that participants who had been tested on blocks one through five recalled approximately 41% of the studied words on the final cumulative test, whereas participants who had been tested only on block five recalled approximately 24% of the studied words, $t(57) = 5.11$, $d = 1.34$, showing a robust testing effect. However, the proportion of critical words produced did not differ significantly between conditions.

TABLE 1
Experiment 1

	Fifth study block test		*Cumulative test*	
	Tested	*Untested*	*Tested*	*Untested*
Studied words*	.50 (0.20)	.36 (0.15)	.41 (0.15)	.24 (.09)
Critical words (of 6)	.01 (0.03)	.09 (0.12)	.17 (0.18)	.21 (.20)
Prior block intrusions	.01 (0.01)	.07 (0.06)	–	–

Mean proportion of studied words, prior block intrusions, and critical words produced on the test after the fifth study block and on the cumulative test for tested and untested participants in Experiment 1 (*SD* in parentheses).
*Refers to fifth study block words for the fifth study block test, and to words from all blocks for the cumulative test.

Discussion

In Experiment 1 we replicated Szpunar et al.'s (2008) finding that interpolated testing counteracts the negative effects of proactive interference by increasing correct recall and reducing prior block intrusions. In addition, testing appeared to reduce false recall: Untested participants produced more critical words on the fifth study block test than did tested participants. However, if the tests taken after each study block are considered, participants in the tested condition produced as many critical words as participants in the untested condition produced on the fifth study block test. This suggests that the activation of the critical words across the five study blocks was equivalent in the two conditions, and not affected by the tests taken after each study block for tested participants. This result makes sense if we consider that activated critical words have, essentially, been ''studied'' (albeit internally). Since participants in the untested condition are unable to efficiently constrain their search set to the fifth study block (due to PI), words that they recall from previous blocks will be produced, and this can include critical words. Furthermore, on the cumulative test the number of critical words produced did not differ between the two groups. Note that there was an increase in false recall from the fifth study block test to the cumulative test (see Marsh & Hicks, 2001 for a similar result); test-induced priming (Marsh, McDermott, & Roediger, 2004) may be responsible for inflating false recall when participants have to produce words from all studied blocks.

The evidence with regards to the effects of testing on false memory that we reviewed in the introduction (Chan & Langley, 2010; McDermott, 1996, 2006) seems to indicate that initial testing may incur the cost of inflating false memories on a later test. We found no such evidence. Neither did we find the opposite pattern: Testing study blocks one through four did not decrease the total number of critical words produced on the cumulative test. Thus testing improved veridical memory and

decreased PI but did not have an effect on false memory. The results so far seem to indicate that testing after every study block improves the ability to distinguish between words that were presented in the most recent block from words presented in previous blocks, but does not necessarily decrease the build-up of false memories. However, since tested participants produced critical words on the test after blocks one through four, they would be unlikely to produce them again on the fifth block test, resulting in the lower level of critical words produced on the fifth study block test compared to untested participants.

EXPERIMENT 2

In Experiment 1 DRM lists were split across the five study blocks to examine how interpolated testing affected the build-up of associative memory illusions from block to block. A potential problem in interpreting Experiment 1 is that false recall of critical words on the initial block tests (but not on the cumulative test) was near floor. This was probably a result of mixing DRM lists in each study block, which has been shown to reduce false recall (Brainerd, Payne, Wright, & Reyna, 2003; McDermott, 1996, Exp. 2; Toglia, Neuschatz, & Goodwin, 1999). In Experiment 2 each study block was composed of words from one single DRM list (for a total of four DRM lists), except for the last study block which was composed of the remaining words from each of the four previously presented DRM lists. So, study blocks one to four were 12-word DRM lists, and the fifth study block consisted of the three remaining words from each of the four presented DRM lists (see Figure 2). With this design we were able to investigate whether having the opportunity to output critical words during initial tests has an impact on the number of critical words produced on the fifth block test where each critical word would be activated, and also on the cumulative test. We discuss predictions for the effects of prior testing on each of the two tests (fifth block test and cumulative test) in turn.

With regards to the fifth block test, based on the results of Experiment 1 we expected participants in the untested condition to produce more critical words on this test because these words would have been activated during study blocks one through four without the chance of being output. As a result, these words would form part of the search set and may be output on the fifth block test, whereas participants in the tested condition would have already had the chance to output each critical word on a previous test and would thus be able to monitor their output to exclude these words. However, taking all five initial study block tests into account, we expected participants in the tested condition to output a larger proportion of critical items than participants in the untested condition would output on the fifth study block test alone, which was not the case in Experiment 1.

With regard to the cumulative test, given the increased number of critical words output on the initial study block tests it is possible that participants in the tested condition would output a higher proportion of critical words on the cumulative test, thus exhibiting a negative consequence of testing. In other words, in this procedure we can assume that the critical words were activated to the same extent during the presentation of each block for every participant, but only participants tested after every block had the opportunity to output a critical word after each studied list. We were thus able to investigate the effect of this initial false recall opportunity on later recall, as compared with the untested condition in which critical words could have been activated but not output. The goal of Experiment 2 was to determine whether blocked DRM lists would produce the same data pattern as mixed DRM lists (with testing producing only a benefit to true recall with no associated cost in terms of increased false recall).

Method

Participants. A total of 35 Washington University undergraduates participated in the experiment for course credit or payment. The participants were randomly assigned to one of two experimental conditions by the program used to run the experiment, resulting in 16 participants in the untested group and 19 in the tested group.

Materials. Four DRM lists were selected from Stadler and colleagues' (1999) norms. The chosen lists were the ones whose critical words were *mountain*, *soft*, *river*, and *city*. From each DRM list the words in the third, fifth, and fourteenth positions were taken and used to form a new study list. These words were randomised within the newly formed fifth study block. So, study blocks one to four were four DRM lists, composed of

12 words each, and study block five was composed of three words from each one of the four DRM lists, also with a total of 12 words. The order of study blocks one to four was counterbalanced across participants, but study block five was always presented last. The words within each block were always presented in the same order for every participant, with exception of the fifth block, which was randomised afresh for each participant (see Figure 2 for a schematic summarising all these details).

Design and procedure. As in Experiment 1 we used a between-participants design with two testing conditions (tested and untested). All dependent measures from the fifth study block test and the cumulative test were identical to those examined in Experiment 1. The procedure and instructions were identical to those in Experiment 1, except that after the immediate cumulative test participants in both testing conditions performed a list discrimination test. This test did not produce differences between the tested and untested conditions, so the results are omitted from this paper for the sake of brevity.

Results

Recall results are presented in terms of proportions as for Experiment 1. The denominators differed to those of Experiment 1 and were as follows: For the fifth study block test: correct recall out of 12 (the number of words presented in the fifth study block); critical words out of 4 (the number of DRM lists used in the experiment); prior block intrusions out of 60 (the number of words presented in study blocks one through four). For the cumulative test: correct recall out of 60 (the total number of words presented across the five study blocks); and critical words out of 4 (the number of DRM lists used in the experiment).

Fifth study block test. The mean proportion of studied words and critical words produced on the fifth study block test in each condition are presented in Table 2. Overall there were no differences between conditions in the proportion of words of each type recalled. More specifically, participants who had been tested after every study block did not correctly recall significantly more words than did participants who were only tested after the fifth study block (the mean proportion of words recalled across the two conditions was .46). The production of critical words also did not differ significantly between participants, with an overall mean across conditions of .05. In fact 3 of the 19 tested participants produced one critical word of the four possible; and 4 of the 16 untested participants produced one of the four possible critical words. However, and contrary to Experiment 1, the mean proportion of critical words produced by tested participants across all five study block tests ($M = .34$) was significantly higher than the mean proportion of critical words produced on the fifth block test by untested participants ($M = .34$ vs .06 critical words recalled on study block five in the untested group); $t(33) = 2.87$, $d = 1.01$. Although the proportion of prior block intrusions was numerically higher for untested participants, this difference did not reach significance.

Cumulative test. The mean proportion of list words and critical words produced on the cumulative test in each testing condition are also presented in Table 2. Participants who had been tested on every study block recalled approximately 53% of the 60 studied words in the final cumulative test whereas participants who had been tested only on list 5 recalled approximately

TABLE 2
Experiment 2

	Fifth study block test		*Cumulative test*	
	Tested	*Untested*	*Tested*	*Untested*
Studied words*	.46 (0.19)	.46 (0.16)	.53 (0.12)	.43 (0.15)
Critical words (of 4)	.04 (0.09)	.06 (0.11)	.33 (0.37)	.36 (0.33)
Prior block intrusions	.01 (0.02)	.04 (0.09)	–	–

Mean proportion of studied words, prior block intrusions, and critical words produced on the test after the fifth study block and on the cumulative test for tested and untested participants in Experiment 2.
*Refers to fifth study block words for the fifth study block test, and to words from all blocks for the cumulative test.

43% of the 60 studied words, $t(33) = -2.17$, $d = 0.76$, replicating the testing effect obtained in Experiment 1. However, the proportion of critical words produced was not significantly different between conditions, with an overall mean of .35.

DISCUSSION

As in Experiment 1, in Experiment 2 we found that tested participants correctly recalled more studied words than untested participants on the cumulative test (due to the testing effect) without an accompanying increase in false recall. Contrary to Experiment 1, in Experiment 2 we did not find any build-up of PI across study blocks, so performance on the fifth block test did not differ between conditions (both in terms of correct recall and prior block intrusions). This most likely occurred as a result of the context change produced by shifting semantic categories between study blocks, as Loess (1968) and Wickens (1970) showed that semantic shift was an effective way of eliminating proactive interference. In addition, unlike in Experiment 1 where untested participants produced more critical words on the fifth block test than did tested participants, in this experiment there were no differences between conditions in the number of critical words produced on the fifth block test. In fact, tested participants actually produced more critical words across blocks one to five than did untested participants on the fifth block test (recall that in the equivalent comparison for Experiment 1, the two numbers were equal). In spite of this, the number of critical words falsely recalled on the cumulative test was equivalent between the two conditions, similarly to Experiment 1.

Overall, Experiments 1 and 2 point to a benefit of initial testing on later veridical memory without the associated cost of false memory. In Experiment 1 previously tested participants performed better on the test after the fifth study block, recalling more presented words and intruding fewer prior-blocks words. These tested participants also falsely recalled less critical words on the fifth block test, but that difference between tested and untested participants did not persist when critical words produced on previous block tests by participants who took those tests were also taken into account. These results point to the effectiveness of testing in increasing veridical memory and protecting against PI (as demonstrated by Szpunar et al., 2008), without the cost of also increasing false recall.

In Experiment 2 previously tested participants did not perform better on the fifth block test than participants in the untested condition, because the context change arising from switching DRM lists between blocks afforded even untested participants a release from PI. Also contrary to Experiment 1, there were no differences in the proportion of critical words recalled in the two conditions on the fifth block test (although both were near floor). On the other hand, when all initial tests were taken into account for the tested condition, participants in this condition produced a higher proportion of critical words than did participants in the untested condition on the fifth block test alone. In spite of this, on the cumulative test participants in the tested condition correctly recalled more studied words, but did not produce a higher proportion of critical words than participants in the untested condition. Across two different sets of materials, and with divergent initial test results, our data suggest that testing improves veridical memory in the long term without the cost of increasing false memories.

EXPERIMENT 3

In Experiment 1 DRM lists were split across the five study blocks so that activation of the critical words could build up from block to block. In Experiment 2, on the other hand, each study block was composed of words from one single DRM list, except for the last study block which was composed of words from each of the four previously presented DRM lists in order to reactivate all four critical words. So Experiment 1 allowed us to study the effects of testing on the build-up of false memories across study blocks, whereas Experiment 2 was specially designed to study the impact of outputting critical words during initial tests on a final cumulative test. However, the levels of false recall on the fifth study block test were rather low in both experiments. In Experiment 1 we already acknowledged that this was most likely the case because words from multiple DRM lists were intermixed rather than being presented in a blocked fashion, which has been shown to reduce false recall of the critical words (Brainerd et al., 2003; McDermott, 1996; Toglia et al., 1999). We attempted to overcome this problem in Experiment 2 by blocking

DRM lists for study blocks one to four, but had to return to intermixed lists in the fifth study block in order to re-activate the critical words and thus be able to compare production of those words in the tested and untested conditions.

In Experiment 3 we presented one DRM list at a time throughout the whole task rather than intermixing words from different DRM lists, even on the fifth study block. To do this, we decreased the length of each study block from 15 to 3 items, so each DRM list was split into five individual study blocks of 3 words each.[3] We then repeated the standard procedure used in previous experiments for each DRM list. That is, for each DRM list participants were either tested on each of the five 3-word study blocks or were only tested on study block five (see Figure 3 for a schematic of one 5-block cycle). This procedure was repeated for all six DRM lists, to measure critical word activation on the fifth study block for every DRM list without intermixing the lists. The goal of Experiment 3 was thus to replicate the results of the previous experiments, with a design that closely mirrored the standard Roediger and McDermott (1995) procedure with the exception of breaking up the DRM lists into short study blocks. That is, at no point in this experiment did participants study intermixed DRM lists.

Method

Participants. A total of 30 Washington University undergraduates participated in the experiment for course credit or payment. The participants were randomly assigned to one of two experimental conditions, resulting in 15 participants in the untested condition (always tested only on the fifth study block of each DRM list) and 15 in the test tested condition (always tested on all five study blocks of each DRM list).

Materials. The same six DRM lists used in Experiment 1 were used in this experiment. The lists were presented in a random order for each participant. Each list was composed of 15 words and was divided randomly into five study blocks of 3 words each. The words within each study block were the same for all participants but presented in a randomised order, and the order of the five study blocks within each list was counterbalanced (see Figure 3 for a schematic with notes on these details).

Design and procedure. The design was a between-participants one, with two conditions (tested and untested). The critical comparison between conditions was recall on the fifth study block tests (one for each DRM list, for a total of six) and on the cumulative test. For the fifth study block tests the level of correct recall, critical word intrusions, and prior block intrusions was measured. For the cumulative test the level of correct recall and critical word intrusions was measured.

The procedure and instructions were similar to those of previous experiments, with the following differences. Each of the six DRM lists was split into five study blocks of three words. After each study block was presented, all participants performed maths problems for 15 seconds. Participants then either solved further maths problems (untested condition), or took a free recall test on that study block (tested condition). The time given for participants to perform the corresponding task after each study block was 15 seconds. All the participants took a test after the fifth study block. On the free recall test after each study block, participants were presented with three boxes and could only enter up to three words. Since participants would be aware that each study block consisted of three words, we wanted to constrain recall so that participants would have to choose which words to report and it was easier for them to understand that they were being tested only on the three words just presented. After all the DRM lists were presented and participants finished the test on the fifth study block of the sixth DRM list, participants were asked to recall for 5 minutes as many words as they could from all of the three-word blocks they had studied.

Results

Recall results are presented in terms of proportions of words recalled of each type. For the fifth study block tests, results are collapsed across the six tests participants took (one for each DRM list). The denominators were as follows. For the fifth study block test: correct recall out of 18 (the total number of words presented in the fifth study blocks across all six DRM lists, i.e., 3×6); critical words out of 6 (the number of DRM lists used in the experiment); prior-list intrusions out of 72 (the number of words presented in study blocks one

[3] We thank Jeff Karpicke for suggesting this experimental design.

through four across all 6 DRM lists, i.e., 12×6). For the cumulative test: correct recall out of 90 (the total number of words presented across the five study blocks); and critical words out of 6 (the number of DRM lists used in the experiment).

Fifth study block tests. The proportion of studied words, critical words, and prior block intrusions produced on the fifth block tests in each condition are presented in Table 3. The mean proportion of words correctly recalled from the fifth study blocks was significantly higher for tested participants than for untested participants, $t(28) = 2.37$, $d = 0.87$. On the other hand, the mean proportion of critical words recalled across all the fifth study block tests was not significantly different between conditions, with an overall mean of .02. In fact, 3 of the 15 tested participants produced one critical word of the six possible; and 4 of the 15 untested participants produced one of the six possible critical words. The mean proportion of critical words produced across all study block tests for tested participants was marginally significantly higher than the mean proportion of critical words produced on the fifth study block tests in the untested group ($M = .13$ vs $M = .03$), $t(28) = 1.62$, $d = .59$, $p = .06$, one-tailed test.

The mean proportion of prior block intrusions was also examined. Tested participants intruded fewer prior block words on the fifth study block tests than did untested participants, $t(28) = 3.53$, $d = .90$.

Cumulative test. The mean proportions of list words and critical words produced on the cumulative test in each condition are also presented in Table 3. Tested participants recalled approximately 38% of the studied words on the final cumulative test, whereas untested participants recalled approximately 31% of the studied words. This difference was significant on a one-tailed t-test, $t(28) = 1.80$, $d = 0.66$ (the results of Experiments 1 and 2 justified the use of a one-tailed t-test for this comparison). However, and also replicating Experiments 1 and 2, the proportion of critical words produced did not differ between conditions, with an overall mean of .31.

TABLE 3
Experiment 3

	*Fifth study block tests**		*Cumulative test*	
	Tested	*Untested*	*Tested*	*Untested*
Studied words**	.86 (0.11)	.72 (0.20)	.38 (0.11)	.31 (0.12)
Critical words (of 6)	.01 (0.04)	.03 (0.09)	.28 (0.33)	.33 (0.24)
Prior block intrusions	.01 (0.01)	.03 (0.03)	–	–

Mean proportion of studied words, prior block intrusions, and critical words produced on the test after the fifth study block and on the cumulative test for tested and untested participants in Experiment 3.
*Results are presented averaged across the six DRM list cycles, each of which consisted of five study blocks.
**Refers to fifth study block words for the fifth study block test, and to words from all blocks for the cumulative test.

Discussion

In Experiment 3 we used a procedure that did not involve intermixing DRM lists at any point during study, to more closely match Roediger and McDermott's (1995) original procedure. On the fifth study block tests previously tested participants recalled more presented words and intruded fewer prior-block words than untested participants, consistent with Szpunar et al. (2008) and our Experiments 1 and 2. Also consistent with the previous experiments, there were no differences between conditions in the number of critical words produced on the fifth study block tests. However, we did successfully detect a significant difference between conditions in prior block intrusions, which were similarly infrequent. Finally, on the cumulative test, tested participants recalled more studied words than untested participants, but the number of falsely recalled critical words did not differ between conditions, once again replicating Experiments 1 and 2. Taken together, these results point to a benefit of testing in preventing the build-up of PI across lists, and increasing the level of veridical memory without the costs of increasing false recall.

GENERAL DISCUSSION

In Experiments 1 and 3 we successfully replicated Szpunar et al.'s (2008) finding that testing helps protect against the build-up of PI with a different set of materials (DRM lists converging on associates). That is, in two different versions of the procedure (intermixed DRM lists in Experiment 1 and pure DRM lists split into three-word study blocks in Experiment 3) we found that participants who were tested on every block produced

more correct words and fewer prior block intrusions on the fifth study block test. In a third experiment (Experiment 2, where we presented blocked DRM lists with a final intermixed study block) we did not observe any prior list intrusions in the untested condition and thus did not find differences in PI between the tested and untested conditions. In addition we replicated the beneficial effect of testing on later recall (e.g., Wheeler & Roediger, 1992) in all three experiments, with tested participants consistently performing better than untested participants on the cumulative recall test.

In addition to replicating Szpunar et al. (2008) with new materials, our design allowed us to look at the effect of interpolated testing on false recall, which to our knowledge has never been examined in a design where each list consists of new words and produces additional activation of the critical words. One potential downfall of testing is that it may increase false recall at the same time as improving veridical recall (e.g., McDermott, 2006). In three experiments we did not find evidence for this pitfall. Below we discuss each of the designs we used and the subtle differences between them; but the take-home point from our data is that testing improved veridical recall and protected against PI without the associated cost of increasing false recall.

In Experiment 1 participants studied five blocks of intermixed DRM lists and were either tested after every block or only after the fifth block. Taking a test after study blocks one to four helped participants to suppress false recall on the fifth block. This could have arisen because for participants in the untested condition, critical words were activated during study of blocks one to four but could not be output, so they formed part of the search set on the fifth study block test. Supporting evidence for this explanation is that tested participants actually produced the same number of critical words across all five study block tests as untested participants produced on the fifth study block test, suggesting that overall activation of the critical words across study blocks was equivalent between conditions. Analogously, on the cumulative test participants in the two experimental conditions produced the same number of critical words.

In Experiment 2 we changed the nature of the study blocks, now presenting participants with pure DRM lists for the first four study blocks and an intermixed list for the fifth study block, in order to re-activate all four critical words. This procedure was a compromise between Roediger and McDermott (1995) who presented blocked DRM lists to achieve high levels of false recall, and the requirement of our design that all critical words were activated following study of the fifth block. Contrary to Experiment 1 there were no differences in PI for tested and untested participants. That is, participants in the untested condition performed as well as participants in the tested condition on the fifth study block test. This absence of PI can be explained by the semantic shift that occurred from study block to study block due to the use of blocked DRM lists, contrary to Experiment 1 where intermixed DRM lists were used in all blocks. We discuss these results below in more detail, but first we turn to the false recall data. The pattern of results for false recall was also somewhat different from that of Experiment 1. In Experiment 2 there was no difference in the number of critical words produced on the fifth study block test between the two experimental conditions (contrary to Experiment 1 where untested participants produced more critical words), but when all study block tests were included for tested participants, the number of critical words produced was higher for tested than untested participants (contrary to Experiment 1 where these numbers were equivalent). However, despite producing more critical words on the initial tests, tested participants produced the same number of critical words as untested participants on the cumulative test, replicating Experiment 1.

In Experiment 3 we did not intermix DRM lists in each study block. Instead participants studied words from one entire DRM list before moving on to the next, just as in the original DRM experiments (Roediger & McDermott, 1995). The difference between the standard procedure and ours was that we presented DRM lists in blocks of three words at a time, with an interpolated task that differed between conditions for the first four study blocks of each DRM list: Untested participants performed 30 seconds of maths problems, whereas tested participants performed 15 seconds of maths problems followed by free recall of the three-word study block. All participants took the free recall test on the fifth study block of each DRM list (for a total of six cycles). Collapsing the results of all the fifth study blocks, we found that participants in the tested condition recalled more words from the correct study block and intruded fewer prior block words; however, there were no differences

in the production of critical words between conditions. The correct recall and prior block intrusion data of this experiment replicate those of Experiment 1, showing that PI built up across study blocks for the untested condition and testing helped relieve this PI. It is worthwhile to note that this is the first demonstration of the Szpunar et al. (2008) in a design with more than one cycle through the procedure. The critical word data of Experiment 3, on the other hand, were more consistent with those of Experiment 2, showing no difference between conditions on the fifth study block tests. However, when comparing the proportion of critical words produced across all tests by tested participants to the proportion of critical words produced on the tests after the fifth study blocks by untested participants, the data seem more consistent with Experiment 1, showing no differences between conditions. Despite these somewhat mixed results, the cumulative test data were clear: participants in the tested condition recalled more studied words and no more critical words than untested participants, replicating both Experiments 1 and 2.

It should be noted that our observed levels of false recall in all three experiments were very low because of the relatively small number of critical words, thus false recall was close to floor on the initial tests. We did find relatively robust levels of false recall on the cumulative test, when the data are considered in terms of proportions. However, these proportions still translate into small numbers of critical words due to the nature of the design, which afforded a maximum of six (Experiments 1 and 3) or four (Experiment 2) critical words to be produced. A fruitful avenue of future research would involve adapting the paradigm to increase the opportunity for false recall on each test, in order to more definitely establish the impact of testing on false recall of related information.

By using lists of words created specifically to elicit false recall of associated words, we were able to measure intrusions due to indirect activation of semantic associates. The results of the experiments point to the conclusion that testing in this paradigm increases true recall without a long-term effect on false recall. This increase in veridical memory unaccompanied by an increase in false memory may seem somewhat at odds with McDermott (2006), Experiment 2), who showed an increase in both veridical and false recall on a cumulative test as a result of initial testing. However, as we noted in the introduction, there was little difference in McDermott's data between the probability of recalling a non-presented critical word after one test (.35) and after three tests (.37); the main effect of testing on critical word recall was most likely driven by the difference between the condition in which no initial test was taken (where the probability of false recall was .27) and the other two conditions. Our untested condition, where participants were tested only after the fifth study block, is more comparable to McDermott's ''one test'' condition than her ''no test'' condition, because in our paradigm the untested participants did take one test (on the fifth study block) prior to the cumulative test. So taking a closer look at McDermott's results demonstrates that our findings are consistent with those reported in that study in demonstrating that increasing the number of prior tests can improve true recall on a final test without an accompanying increase in false recall. It should be noted that previous work on the effects of testing on false recall has focused on repeated recall of the same information. We adopted a new procedure because our goal was to assess the effects of testing different but related material on veridical memory, proactive interference, and false memory for semantically related information.

One of the explanations of the benefits of testing suggested by Szpunar and colleagues (2008) was the source monitoring/reduction of cue overload explanation. This explanation seems to be supported by our results. According to this view, if participants take a test after each study block, they develop a set of temporal cues associated with each retrieval moment (for each of the tests after blocks one through four); so, on the fifth block test, the words associated with the effective retrieval cue at the time of testing will only be words from the fifth block, since the previously studied words will have already been associated with previous retrieval cues. Therefore participants who have been tested after each study block can easily circumscribe the relevant search set to words presented in the most recently studied block. On the other hand, participants who have only studied (and not been tested on) previous blocks will not have access to such cues because all previously presented words (those from study blocks one through four, and those from study block five) will have been associated with the same cue. This prevents participants from being able to circumscribe the relevant search set to the items studied only in the most recent block. As a result, the level of correct recall of words from that block

decreases, as more retrieval candidates are subsumed under the retrieval cue for that test (Watkins & Watkins, 1975). Interference from previously presented words and indirectly activated critical words (which could also have been activated during study of previous blocks) increases because all of those words share the same retrieval cue (Szpunar et al., 2008).

However, on the cumulative test the retrieval cue should be the same for all participants, no matter whatever they were tested after each study block or only after the fifth. In this case the relevant search set will include words presented in all five study blocks as well as indirectly activated critical words (or, if such words were produced on the initial block tests, directly activated critical words). As the words share the same retrieval cue, the temporal cues will no longer be helpful in differentiating presented words from non-presented words, because those could have been activated at any point during study.

The cue overload explanation is also supported by the finding that the act of recalling from long-term memory may drive context change, isolating the recalled list of words from interference. Jang and Huber (2008) used a retroactive interference paradigm in which participants are asked to recall not from the most recently presented list, but from a previously presented one (the target list). Jang and Huber manipulated the length of both the target list and the intervening lists, and whether participants received a free recall test after every intervening list and the target list, or only after the target list. Only the length of the target list affected free recall of the target list when participants were tested on all intervening lists. On the other hand, both the length of the target list and the length of the intervening lists affected recall of the target list when participants were not tested on the intervening lists. These results are interpreted as supporting the view that the act of recall drives context change, isolating the target list from interference from intervening lists. Further support for the context change explanation was put forward by Pastötter et al. (2011), who found that a semantic generation task and a working memory task both protected against PI similarly to intervening testing.

Although PI was not the focus of our paper, we found an interesting difference between our experiments in terms of PI. In Experiments 1 and 3 we found the expected difference in the learning of study block five between the previously tested and untested conditions. A context change afforded by the act of retrieval (Jang & Huber, 2008; Pastötter et al., 2011) may explain why testing protects against proactive interference, as it provides a change similar to the one that occurs when the category of the list is changed (e.g., Wickens, 1970). However, this difference did not emerge in Experiment 2 because PI was minimal even for those who had not been tested on study blocks one through four. It is interesting to note that this absence of PI emerged despite the fifth block being semantically related to all of the previous blocks. It seems that the change in semantic context between blocks one through four was enough to keep PI at bay for untested participants, even though there was some semantic overlap between block five and all four previously studied blocks. This release from PI due to a change in semantic dimension or context is highly congruent with classic studies showing release from PI. For example, Wickens (1970) showed that semantic dimensions, including taxonomic categories, are highly effective in promoting release from PI; and Wickens, Dalezman, and Eggemeier (1976) showed that the effect of semantic shift on release from PI was negatively correlated with the property overlap between semantic categories (e.g., more release occurs when switching from fruits to occupations than from fruits to vegetables). It is therefore easy to see why we did not find PI in our Experiment 2. Although we did not anticipate this effect, it gave us the chance to demonstrate equivalent effects of testing on false recall in two situations: One in which testing protect against the build-up of PI across study lists (Experiments 1 and 3) and another in which PI did not build up across lists because of semantic shifts between lists (Experiment 2).

Our results can shed light on a recently proposed theoretical explanation for the testing effect. Carpenter (2009) proposed the *elaborative retrieval hypothesis*, which states that retrieval processes during testing may activate elaborative information related to the studied item. This elaborative information may then help later retrieval, by providing additional cues to the target item. This hypothesis was mainly developed to explain the benefits of testing in a paired associate cued recall paradigm, and does not seem consistent with our results that testing improves veridical recall without increasing false recall. According to the elaborative retrieval hypothesis, critical non-presented words should receive extra activation during the tests on study blocks one to

four and thus they should be falsely recalled more often on the final cumulative test by participants who had been tested after every study block. This was not the case in any of our three experiments. Of course, the elaborative retrieval hypothesis may be applicable to cued recall tests performed after the study of pairs of words and it is compatible with the notion that tests enhance retention because they promote encoding of mediating information—i.e., a word or concept that links a cue to a target, as demonstrated by Pyc and Rawson (2010). However, according to our results, the elaborative retrieval hypothesis does not seem to be applicable to the testing effect observed in free recall.

In conclusion, we investigated a potential negative effect of testing on long-term memory: An increase in false memories due to increased activation of critical words during retrieval practice. The existence of this negative effect of testing was not supported by our results. Instead, we found that testing improved performance on a cumulative free recall test by enhancing true recall through retrieval practice (Experiments 1, 2, and 3), and improving learning by protecting against PI throughout the study of multiple consecutive lists (Experiments 1 and 3). On the other hand, testing did not affect false recall on the final cumulative test, even when a higher number of critical items were produced on the initial tests by tested participants (Experiment 2). Our results show a dissociation between true and false recall, and demonstrate that it is possible to increase the correct recall of studied information without necessarily increasing the false recall of associated information. In our study veridical recall was improved by testing without the cost of increasing false recall of associated information.

Manuscript received 19 September 2011
Manuscript accepted 6 December 2011
First published online 25 January 2012

REFERENCES

Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language*, *48*, 445–467.

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

Chan, J. C. K., & Langley, M. (2010). Paradoxical effects of testing: Retrieval enhances both accurate recall and suggestibility in eyewitnesses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 248–255.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic memory. *Psychological Review*, *82*, 407–428.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusion in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

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, & Cognition*, *34*, 112–127.

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

Johnson, M. K., & Raye, L. C. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Karpicke, J. D., & Roediger, H. L. III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning & Verbal Behavior*, *1*, 153–161.

Loess, H. (1968). Short-term memory and item similarity. *Journal of Verbal Learning & Verbal Behavior*, *8*, 240–247.

Marsh, R. L., & Hicks, J. L. (2001). Output monitoring tests reveal false memories of memories that never existed. *Memory*, *9*, 39–51.

Marsh, E., McDermott, K. B., & Roediger, H. L. (2004). Does test-induced priming play a role in the creation of false memories? *Memory*, *12*, 44–55.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory & Language*, *35*, 212–230.

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

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

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General*, *106*, 376–403.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *333*, 335.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*, 231–237.

Roediger, H. L., & Gallo, D. A. (2004). Associative memory illusions. In R. F. Pohl (Ed.), *Cognitive illusions*. New York, NY: Psychology Press.

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

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications

for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 803–814.

Roediger, H. L., McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in producing false remembering. In M. Conway, S. Gathercole, & C. Cornoldi (Eds.), *Theories of memory II* (pp. 187–245). Hove, UK: Psychology Press.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407.

Stadler, M. A., Roediger, H. L. III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *27*, 494–500.

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

Toglia, M. P., Neuschatz, J. S., & Goodwin, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, *7*, 233–256.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*, 49–60.

Watkins, O. C., & Watkins, M. J. (1975). Build-up of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *104*, 442–452.

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

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

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to word meaning. *Psychological Review*, *77*, 1–15.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in shortterm memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 440–445.

Wickens, D. D., Dalezman, R. E., & Eggemeier, F. T. (1976). Multiple encoding of word attributes in memory. *Memory & Cognition*, *4*, 307–310.