Journal of Experimental Psychology:
Learning, Memory, and Cognition
2014, Vol. 40, No. 4, 1039–1048

0278-7393/14/$12.00 DOI: 10.1037/a0036164

The Role of Test Expectancy in the Build-Up of Proactive Interference in Long-Term Memory

Yana Weinstein
University of Massachusetts–Lowell

Adrian W. Gilmore
Washington University in St. Louis

Karl K. Szpunar
Harvard University

Kathleen B. McDermott
Washington University in St. Louis

We examined the hypothesis that interpolated testing in a multiple list paradigm protects against proactive interference by sustaining test expectancy during encoding. In both experiments, recall on the last of 5 word lists was compared between 4 conditions: a tested group who had taken tests on all previous lists, an untested group who had not taken any tests on previous lists, and 2 other groups (one tested and the other untested) who were warned about the upcoming test prior to study of the fifth list. In both experiments, the untested/warned group performed significantly better than the untested/unwarned group on both correct recall and prior list intrusions but did not achieve the same recall accuracy as tested groups. In Experiment 2, an instruction manipulation check further narrowed the gap between the untested/warned group and the tested groups. In addition, we verified that a reduction in test expectancy indeed occurred in the untested group compared with the tested group by asking participants to indicate how likely they believed they were to receive a test on each studied list. These findings suggest that testing protects against proactive interference largely via attentional processes and/or more effective encoding.

Keywords: proactive interference, interpolated testing, test expectancy, encoding

Supplemental materials: http://dx.doi.org/10.1037/a0036164.supp

Testing has been shown to benefit later retrieval (Gates, 1917; Karpicke & Roediger, 2007), possibly because the act of retrieval itself strengthens that information in memory (e.g., Inda, Muravieva, & Alberini, 2011). Szpunar, McDermott, and Roediger (2008; see also Bäuml & Kliegl, 2013; Weinstein, McDermott, & Szpunar, 2011) further demonstrated that testing can help strengthen not only practiced information but also new information studied after the test is taken. When multiple sets of materials are studied, taking a test after each set and before studying the next set(s) appears to help participants learn the later material better. This finding has led researchers (e.g., Szpunar et al., 2008) to hypothesize that taking interim tests greatly reduces the proactive interference (PI) that normally accumulates during extended bouts of studying (see also Postman & Keppel, 1977).

In Szpunar et al.'s (2008) experiments, participants studied five lists of words. Following the study of each list, participants completed 1 min of math problems and then completed either a further set of math problems or tried to remember the words from the list they had just studied, depending on experimental condition. After the fifth list, participants in both groups were given a test on that list. Participants who had been tested after Lists 1–4 performed better on List 5 than did participants who had continued to do math problems instead of taking the tests. More specifically, correct recall of the words from List 5 was improved and intrusions from Lists 1–4 were reduced.

How does taking a test on previously studied information help when it comes to learning new information? One possibility is that prior tests allow participants to encode subsequent lists in the study sequence as effectively as the first list. Using electroencephalography, Pastötter, Schicker, Niedernhuber, and Bäuml (2011) demonstrated that alpha power during encoding of a list after previous lists have been studied is reduced following an unrelated control task and remains at the same level as that recorded during the first list following a test. Alpha power is a brain oscillation frequency that has been linked to memory load (see Jensen, 2006, for a review) and also inattention (see Palva & Palva, 2007, for a review), suggesting that when alpha power increases, successful encoding is less likely. The authors concluded that retrieval between lists helps to "reset" pro-

This article was published Online First April 7, 2014.

Yana Weinstein, Department of Psychology, University of Massachusetts–Lowell; Adrian W. Gilmore, Department of Psychology, Washington University in St. Louis; Karl K. Szpunar, Department of Psychology, Harvard University; Kathleen B. McDermott, Department of Psychology, Washington University in St. Louis.

Support for this research was provided by a James S. McDonnell Foundation 21st Century Science Initiative grant: Bridging Brain, Mind and Behavior/Collaborative Award.

Correspondence concerning this article should be addressed to Yana Weinstein, Department of Psychology, University of Massachusetts–Lowell, 113 Wilder Street, Suite 300, Lowell, MA 01854. E-mail: Yana_Weinstein@uml.edu

cesses associated with these alpha waves and allows for continuously efficient encoding across multiple lists.

Another possibility is that prior tests help participants discriminate information that comes to mind during subsequent tests. In interpreting their original results, Szpunar et al. (2008) suggested that, relative to previously untested lists, words from previously tested lists are more easily distinguished when encountered later on in the experiment (i.e., participants are better able to recall the source of words from tested lists; Chan & McDermott, 2007). Related to this position, Szpunar et al. (2008) further suggested that testing might serve to create contextual cues that can be used to distinguish between previously recalled words (i.e., words that were studied in previous lists) and words from the list just studied (see also Postman & Keppel, 1977).

Clearly, the extent to which the insulating effect of testing on proactive interference is an effect of encoding, retrieval, or some combination of the two remains to be hashed out. Importantly, there is an additional potential influence that existing accounts have not yet considered: attentional processes and/or quality of encoding may change over time as a result of participant-derived expectations of future testing. That is, participants may be processing information more effectively when they expect imminent tests, and this effectiveness could be a result of either increased attention, or improved quality of encoding, or both. For the rest of the article, we refer to this explanation in terms of attentional processes for brevity, but it should be noted that quality of encoding could also be a candidate process. The experiments we present here were designed to test this additional influence of test expectancy[1] on recall of interpolated word lists.

In Szpunar et al.'s (2008) paradigm, all participants regardless of condition were given a cumulative free recall test after study of all the lists. Participants were told to prepare for this test, so all participants expected a final cumulative test. Moreover, participants were also told to expect tests after some of the individual lists. In particular, they were told that the computer program would determine randomly after each list whether they did or did not get a test. In reality, of course, there were only two testing schedules: either there was a test after every list and then a cumulative test (for participants henceforth referred to as the *tested group*), or there was a test only after the fifth and final list and then a cumulative test (for the *untested group*). Given these different testing schedules, a possible alternative explanation of the insulating effect of testing is that the expectation of an imminent test, rather than the experience of previous tests, could be driving the apparent beneficial effect of testing on the learning of subsequent lists (cf. Szpunar, McDermott, & Roediger, 2007). It is reasonable to assume that participants in the tested group would have been expecting a test after the fifth list, having had a test after Lists 1–4, whereas participants in the untested group may not have been expecting this test. The argument could be made, then, that participants in the untested group would have been paying less attention or engaging in lower quality encoding strategies by the time the fifth list came around, because the likelihood of a test on that list seemed low.

The experiments reported here are based on the assumption that participants in the tested group may be more likely to expect a test after the fifth (last) list, having consistently received tests after previous lists. Those in the untested group, having never received a test during the experiment until the fifth list test, may therefore pay less attention or engage in lower quality encoding strategies during encoding of the fifth list. To test for this alternative explanation, we compared the two standard conditions (tested and untested) with two novel conditions that were identical to the standard conditions but included a warning before presentation of the final list to alert participants that they would be tested on the upcoming list. If attentional processes are mediating the observed release from proactive interference, warning participants in the untested group should produce the same benefit as the participants taking a test after every list. In Experiment 2, we also directly tested the assumption that expectations of a test on List 5 differ between the tested and untested groups. The question we set out to address was whether the benefits of interpolated testing are at least in part driven by the expectation of experiencing further tests.

Experiment 1

Method

Participants. The sample consisted of 250 U.K.-based adults (164 or 65.6% females) ages 18–30 years ($M = 24.7$; $SD = 3.6$) who participated in the experiment online and received loyalty points (redeemable for cash and vouchers) for their participation. Of the 250 participants, 154 (61.6%) had an undergraduate degree or a higher degree or were attending college. Twenty-two participants were not native English speakers, but 16 of them had been speaking English for more than 10 years, and the rest had been speaking English for at least 4 years.

Materials. The materials were identical to those used by Szpunar et al. (2008, Experiment 1A) and consisted of five interrelated study lists of 18 words. Each list included three words from each of the following six semantic categories: building parts, earth formations, animals, fruits, human body parts, and weather phenomena.

Design. This experiment involved a 2×2 between-subjects design. Participants were either tested after every list (tested groups) or tested only after the fifth list (untested groups). Orthogonally, participants were either warned about the test occurring on the fifth list prior to study of that list (warned groups), or they received no such warning (unwarned groups). Figure 1 represents a schematic of the procedure in each of the four conditions. The computer program determined group membership and the order of the study lists randomly for each participant who accessed the online experiment. This resulted in 60 participants in the tested/unwarned group, 65 participants in the untested/unwarned group, 63 participants in the tested/warned group, and 62 participants in the untested/warned group. Correct recall and prior list intrusions on the List 5 test and correct recall on the final cumulative test were compared among the four groups. Each of these measures was subjected to a univariate analysis of variance (ANOVA), and comparisons are significant to $p < .05$ unless otherwise stated.

Procedure. Eligible participants enrolled in the Maximiles online loyalty scheme (see www.maximiles.co.uk) received an e-mail inviting them to participate in an experiment in exchange

[1] While the term *test expectancy* has been used in the literature to denote the expectation of a particular type of test (e.g., recall vs. recognition; Balota & Neeley, 1980), we use it in this article simply to denote the expectation of any kind of test versus no expectation of a test.

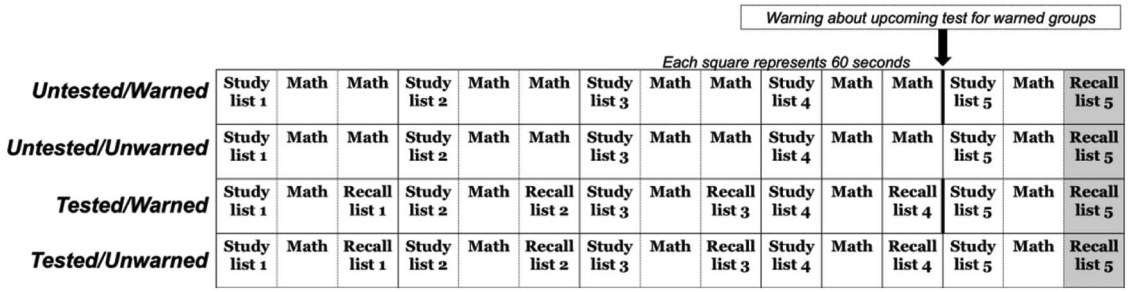

Figure 1. Schematic representing the procedure in each of the four experimental conditions.

for loyalty points. The experiment was programmed in Adobe Flash (Adobe Systems, San Jose, CA) and hosted on the first author's website. Interested individuals followed the link and completed the experiment online. Upon agreeing to the online consent form, participants were told that they would study five lists of words (2,000 ms per word; 500 ms interstimulus interval) and that after the presentation of each individual list, they would complete 1 min of math problems that would be followed by either 1 additional minute of math or a test that would require them to recall the words from the immediately preceding list. In addition, participants were warned that they would be asked to recall all the words from all of the lists on a final cumulative test. Participants were informed that the computer program determined the occurrence of a test after each list randomly. In reality, there were only two testing schedules: (a) testing after every list (tested groups) and (b) testing after the fifth list only (untested groups). Thus, after studying each list and completing 1 min of math problems, participants either spent the next minute recalling as many words as they could remember from that list (tested groups), or completing math problems for an additional minute (untested groups). After the fifth list, all participants took the recall test following 1 min of math problems. For these tests, participants typed the words on the computer as they recalled them; each newly typed word disappeared from view when participants hit the Enter key. Finally, immediately after the fifth list test, participants recalled as many words as they could from all of the lists.

In addition to being randomly assigned to a tested or untested group, participants were also randomly assigned to be in one of two different warning conditions. Unwarned participants performed the task as described in the preceding paragraph. Participants who were assigned to a warned condition received a warning about the upcoming test before studying List 5. The following instructions appeared for the warned groups before presentation of the fifth (i.e., the final) list:

> Regardless of whether you have received tests after any of the previous lists that you have studied, *you WILL get a test after the next list*. That is, just after you have studied the next list and completed 1 min of math problems, you will be asked to recall as many words as you can from that list. So, please pay attention to the following list, as you will be tested on it. Following that test, you will also be asked to recall the words from ALL the lists you have studied.

The phrase italicized in the above instructions was displayed in red font to draw participants' attention to this important message. The word "WARNING" also appeared in large red font above the instructions.

Results

Exclusion criteria. Since the experiment was not carried out in the lab, and there was less control over participants' behavior, two exclusion criteria were used. First, at the end of the experiment, participants were asked whether they noted down any words during the quiz (e.g., on a piece of paper) to help on the memory test. The response options were "I did not note down any words"; "I noted down the occasional word"; "I noted down about half the words"; and "I noted down all of the words." Only participants who indicated that they did not note down any words (205 of 250, or 82.0% of participants) were included in any analyses. Second, a further 13 participants were excluded from the study because they did not appear to engage in the task in that they did not recall any words on the fifth list and/or cumulative tests. These exclusion criteria resulted in 46 participants in the tested/unwarned group, 52 participants in the untested/unwarned group, 41 participants in the tested/warned group, and 53 participants in the untested/warned group.

Initial list tests. Figure 2 shows the number of words correctly recalled on Lists 1–5 for the tested groups and only on List 5 for the untested groups. As can be seen from the figure, correct recall in the tested group declined somewhat across the five lists. These data were submitted to a mixed ANOVA with warning group (warned vs. unwarned) as the between-subjects variable and list number (1–5) as the within-subject variable. Only the within-

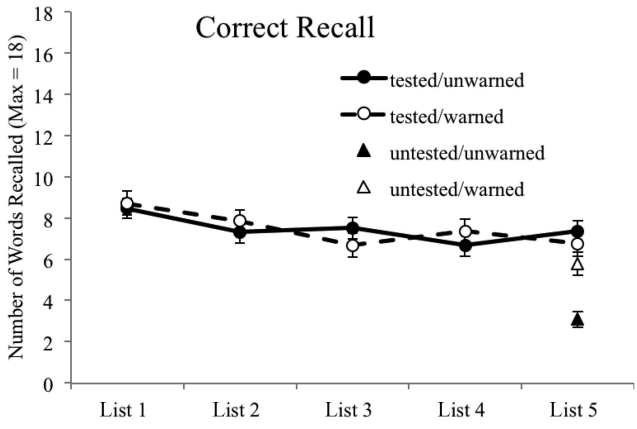

Figure 2. The number of words correctly recalled on each studied list in the two tested groups and on the fifth list in the two untested groups in Experiment 1. Error bars represent standard error of the mean.

subject variable list number had an effect on recall, $F(4, 340) = 6.43$; partial $\eta^2 = .07$, $p < .001$, for the within-subject ANOVA on correct recall across Tests 1–5.

Looking now at List 5 performance across all four conditions, participants who had been tested on Lists 1–4 performed better than the untested groups. Correct recall from the List 5 test alone was subjected to a univariate ANOVA with a 2 (testing group: tested/untested) × 2 (warning group: warned vs. unwarned) design. First, there was a main effect of testing group such that participants in the tested groups performed better on the List 5 test than participants in the untested groups, $F(1, 188) = 24.5$; partial $\eta^2 = .12$, $p < .001$. There was also an effect of warning, such that participants in the warned groups performed better on the List 5 test than participants in the unwarned groups, $F(1, 188) = 4.03$; partial $\eta^2 = .02$, $p = .046$. Crucially, though, both of these main effects were qualified by an interaction between testing and warning, $F(1, 188) = 9.73$; partial $\eta^2 = .05$, $p = .002$. This interaction indicated that when participants had not received tests on Lists 1–4, a warning prior to study of the fifth list almost doubled performance on that test, $t(92.4) = 3.93$, $p < .002$ (corrected for unequal variances). On the other hand, when participants had been consistently tested on Lists 1–4, performance was not improved by the warning. Critically, in a planned comparison, the untested/warned group did perform at a significantly lower level than the tested/unwarned group in terms of correct List 5 recall, $t(97) = 2.04$, $p = .044$.

The data for prior list intrusions are presented in Figure 3 and are directly analogous to the correct recall data. As can be seen from the figure, prior list intrusions were consistently low on Lists 2–5 for the tested groups but did increase steadily across lists; $F(3, 255) = 3.91$; partial $\eta^2 = .04$, $p = .009$, for the mixed ANOVA on prior list intrusions across Tests 2–5. Neither the main effect of warning group nor the interaction between testing and warning was significant, ps $> .127$. As with correct recall, on the fifth list test, there was a main effect of testing condition, $F(1, 188) = 57.1$; partial $\eta^2 = .23$, $p < .001$, a main effect of warning, $F(1, 188) = 13.5$; partial $\eta^2 = .07$, $p < .001$, and an interaction between testing and warning on prior list intrusions, $F(1, 188) = 8.72$; partial $\eta^2 = .04$, $p = .004$. Mirroring the correct recall results, this interaction indicated that when participants had not received tests on Lists 1–4, a warning prior to study of the fifth list halved the number of prior list intrusions on that test, $t(103) = 3.75$, $p < .001$. On the other hand, when participants had been consistently tested on Lists 1–4, prior list intrusions were not significantly reduced by the warning. In a planned critical comparison, though, the untested/warned group did produce more prior list intrusions than the tested/unwarned group, again mirroring the correct recall results, $t(66.5) = 3.29$, $p = .002$.

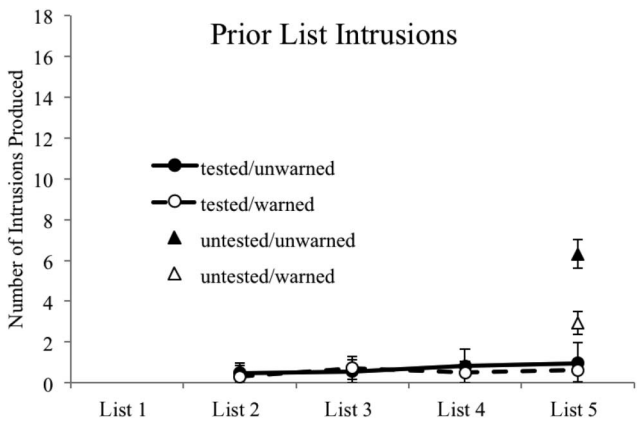

Figure 3. The number of prior lists intrusions produced on Lists 2–5 in the tested groups and on the fifth list in the untested groups in Experiment 1. Error bars represent standard error of the mean.

Cumulative test. The left side of Figure 4 presents the proportion of words recalled correctly on the cumulative test as a function of testing and warning conditions. On this test, there was only a significant main effect of testing condition; $F(1, 188) = 32.8$; partial $\eta^2 = .15$, $p < .001$ (ps $> .494$ for the main effect of warning and the interaction), such that on average, the tested groups produced a mean of 27.2 ($SD = 13.8$) of the 90 words they had studied, whereas the untested groups produced a mean of 16.9 ($SD = 11.3$) of the 90 words. When examining just the recall from List 5 on the cumulative test, as shown on the right side of Figure 4, a different pattern emerged. The significant main effect of testing, $F(1, 188) = 8.24$; partial $\eta^2 = .04$, $p = .005$, was qualified by an interaction between testing and warning, $F(1, 188) = 9.20$; partial $\eta^2 = .05$, $p = .003$. That is, the untested group recalled more words from List 5 on the cumulative test when they had been warned about the test prior to study of that list, although this effect was marginal, $t(98.1) = 1.70$, $p = .092$ (corrected for unequal variances), whereas the effect was in the opposite direction for the tested group, with significantly fewer words recalled from List 5 on the cumulative test by the group who had been warned of the test prior to study of that list, $t(85) = -2.60$, $p = .011$.

Discussion

The purpose of the present study was to investigate whether attentional processes play a role in creating the benefit of testing against proactive interference. The results clearly demonstrate that this, at least in part, appears to be the case. Replicating the findings of Szpunar et al. (2008), we found that, relative to participants who had not been tested after Lists 1–4, participants who were tested after Lists 1–4 recalled more than twice as many words from List 5. This benefit of prior testing was similarly manifested in the intrusion data. Building on these prior findings, we also showed that providing previously nontested participants with a warning (of an impending test) before presentation of List 5 helped them to recall more words from that list and to produce fewer prior list intrusions. Warned but untested participants largely recovered to the level of the tested participants, but it should be noted that they did not quite reach equal performance for either correctly recalled words or intrusions.

An intuitive explanation for the results obtained in Experiment 1 is that participants who are tested after every list expect a test after List 5, whereas participants who are not tested on Lists 1–4 do not and that this expectancy accounts for most of the differences observed so far in the literature. The warning prior to study of List 5 would then narrow the gap in performance between the two conditions by aligning test expectancy. This explanation involves assumptions that can be tested empirically—that is, that test expectancy is equivalent in the two conditions (tested and untested) at the first list, but gradually diverges as participants in the

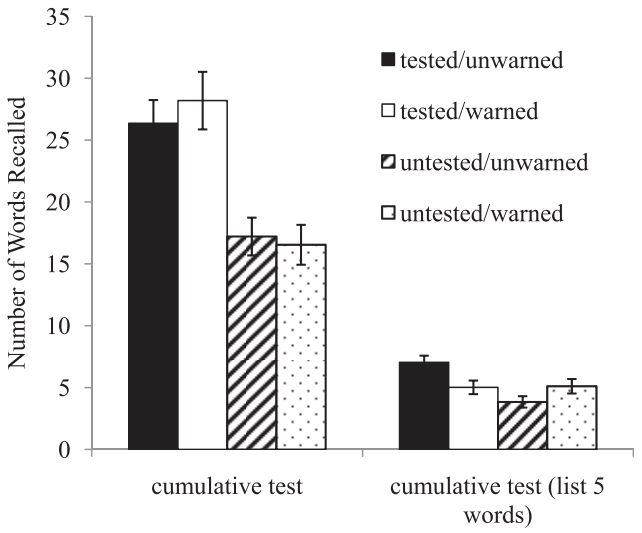

Figure 4. The number of words correctly recalled on the cumulative test in Experiment 1 in each condition, from all studied lists (left side) and from the fifth list only (right side). Error bars represent standard error of the mean.

tested condition keep receiving tests, whereas participants in the untested condition receive no tests. A further assumption is that the warning before List 5 in the warned group serves to increase expectancy to the same level as for those participants who have been receiving tests after every list, whereas the same warning in the tested group would have little or no effect due to test expectancy already being high. In Experiment 2, we tested these assumptions.

Experiment 2

In this experiment, we utilized the same basic paradigm as Experiment 1, but before studying each list, half of the participants were asked to provide a rating of how likely they felt they were to receive a test after studying the list. For the two unwarned conditions, this allowed us to track, across the five studied lists, whether participants in the tested or untested condition felt they were more likely to receive a test after each study period. To the extent that expectancy influences performance on the fifth list test, presumably by affecting attentional processes, it is important to know if participants in the tested and untested groups had different expectations about receiving a test on every list. For the two warned conditions, this also allowed us to measure the effect of the warning on test expectancy. Based on the results of Experiment 1, we predicted that we would see an interaction between testing and warning conditions, such that participants in the tested conditions would have high test expectancy before List 5, whereas only those untested participants who received a warning (compared with the unwarned/untested participants) would have a high test expectancy.

An additional feature of this experiment was an extensive attempt to focus specifically on participants who were actively engaged in reading and processing our instructions. That is, in Experiment 1, we saw that receiving a warning before study of List 5 substantially improved memory of that list for untested participants but not quite to the level of having had prior tests on every list. One potential explanation of these results is that a small subset of participants in the untested/warned group did not read or attend to the warning and thus behaved just like participants in the untested/unwarned group.[2] In Experiment 1, we did not implement any kind of procedure to ensure that participants in the warned group actually read and understood the instruction indicating that they would definitely be tested on the fifth list. This is of particular concern since the experiment was administered online, affording us relatively little control over participants' engagement with the task. In this experiment, we therefore included multiple checks throughout the experiment to ensure that participants were reading and understanding the instructions (see Crump, McDonnell, & Gureckis, 2013, for the effectiveness of such measures in online experiments).

In Experiment 2, we had three aims. First, we wanted to test the hypothesis that participants in the untested condition would effectively stop believing that they would be tested on subsequent lists after experiencing multiple untested lists. Second, we sought to replicate the novel pattern of data we presented in Experiment 1—namely, that presenting a warning prior to study of a list can significantly reduce proactive interference both in terms of increasing correct recall and reducing prior list intrusions. Third, we were interested in whether an instructional manipulation check would further improve release from proactive interference after a warning, in particular because the sample would exclude any participants who missed the warning.

Method

Participants. The sample consisted of 447 U.S.-based adults (230 or 51.5% females) ages 18 to 69 ($M = 32.2$; $SD = 11.1$) recruited from Amazon Mechanical Turk. Of the 447 participants, 406 (90.1%) had an undergraduate degree or a higher degree or were attending college. Twelve participants were not native English speakers; of these, nine had been speaking English for more than 10 years, and the rest had been speaking English for at least 4 years.

Design and procedure. This experiment utilized a $2 \times 2 \times 2$ between-subjects design. Participants were either tested after each list (tested condition) or only tested after the fifth list (untested condition). In addition, participants were either warned about the List 5 test prior to studying the fifth list (warned condition) or were given no such warning (unwarned condition). Finally, participants were either asked to make test expectancy ratings before each list to indicate how likely they felt they were to be tested on that list (rating condition) or were never asked to provide these types of ratings (no rating condition). Crossing these three different factors resulted in eight experimental conditions, and participants were randomly assigned to one of them by the computer program.

The experiment was posted as a HIT (task) on Mechanical Turk. Only Mechanical Turk workers who were located in the United States and had a minimum acceptance rate of 95% on all previous tasks were eligible to participate.

Materials and methods were identical to those of Experiment 1, with the following exceptions. First, participants were instructed to check specific boxes on each of two instruction screens at the

[2] We thank Evan Heit for suggesting this possibility.

beginning of the experiment to indicate that they had read the instructions. The instructions pertaining to which boxes had to be checked appeared in different places for the two instruction pages, so participants had to be reading the instructions carefully in order to comply with them. Figures S1a and S1b in the online supplemental materials show screenshots of these instructions. Second, participants had to summarize the instructions that were given prior to the cumulative final test (i.e., they had to specify that they were to write down words from all lists). After reading the cumulative test instructions, participants were asked to complete the following sentence: "You will now be asked to recall as many words as you can from . . . ." The correct response was some variation of "all five lists." The first two amendments just described pertained to all participants, whereas the third and fourth described later pertained to the warned groups and the rating groups, respectively.

The third methodological change from Experiment 1 was that participants in the warned groups had to reproduce part of the warning that occurred prior to List 5. Immediately after the warning was presented, participants in the warned group had to respond to the following question, "What did the warning on the previous screen say? Please complete the following sentence: 'Just after you have studied the next list and completed 1 min of math problems, you will . . . .'" Participants were given as long as they needed to type a response into the textbox, before moving on to study the fifth list.

The fourth methodological change from Experiment 1 was that participants in the ratings groups had to indicate how likely they felt they were to be tested on that list, prior to studying the subsequent list of words. Specifically, participants were given the following instructions:

> You are about to study a list of words. Before you do so, though, we want to ask one question: Do you think there will be a memory test on this list (before you move on to the next list of words)? Please just take your best guess.
>
> Please move the slider to the appropriate position (indicating whether or not you think there will be a test and your level of confidence about that guess). That is, if you're not at all sure, you would place your slider closer to the middle. If you're sure there will be a test, you'd place it far to the right, and if you're sure there won't be one, you'd place it far to the left.

After moving the slider to indicate their test expectancy for the upcoming list, participants then proceeded to study the words in that current list. Note that participants in the warned/ratings groups were given the List 5 warning prior to reporting their test expectancy ratings. In addition, as in Experiment 1, all participants were told to expect a final cumulative test on all five lists.

Results

Exclusion criteria. Only participants who passed all exclusion criteria were kept for subsequent analysis. First, 80 participants were excluded for noting down words during the experiment, the exclusion criterion that was used in Experiment 1. Second, 15 participants failed the initial instruction check (i.e., failed to demonstrate that they initially read the instructions by clicking on the correct boxes as indicated). Third, eight participants were excluded for failing to reiterate their instructions for the final cumulative recall test. Two additional participants were excluded because they did not appear to engage in the task (i.e., they entered random words or phrases during the List 5 or cumulative recall tests). These exclusion criteria relate to all groups. The fourth exclusion criterion related only to the warned groups, where eight additional participants were excluded for failing to reiterate the warning prior to study of List 5. Finally, seven participants were excluded because they completed the task from the same Internet protocol (IP) address as a previous participant (this could be due to multiple family members using the same computer, but the precaution was taken in case the same participant had completed the task multiple times), and two participants' data were unusable due to a programming error. This led to exclusion of 122 of the 447 total participants (27.3%), which is in line with other studies that draw from the same population and use instructional manipulation checks (e.g., Goodman, Cryder, & Cheema, 2013). Table S1 in the supplemental materials presents exclusion frequencies for each condition by exclusion reason, for archival purposes.

Test likelihood predictions. Mean test likelihood predictions as a function of testing group for each list can be found in Figure 5. It is important to note that at the time the first judgment was made, all groups had had identical experiences, so there was no reason to expect any difference in ratings of test likelihood between any of the conditions on List 1. The primary comparison of interest was for test likelihood predictions prior to study of List 5, and this comparison is described in detail in the next paragraph. However, we also conducted a 4 (list number 2/3/4/5) × 2 (testing group: tested vs. untested) × 2 (warning group: warned vs. unwarned) repeated-measures ANOVA on test expectancy ratings. Looking at Figure 5, we can see that prior to study of the first list, all participants were fairly confident that they might be tested on that list ($M = 72.4$, $SD = 22.7$ across all conditions). It is not until List 3 that a pattern separating the tested versus the untested groups emerged, such that the tested groups remained fairly constant in their expectations of a test following each list, whereas the untested groups lost confidence that they would be tested. The analysis revealed main effects of list number, $F(3, 486) = 14.1$, partial $\eta^2 = .08$, $p < .001$, and a main effect of testing group, $F(1, 162) = 35.8$, partial $\eta^2 = .18$, $p < .001$. In addition, there were

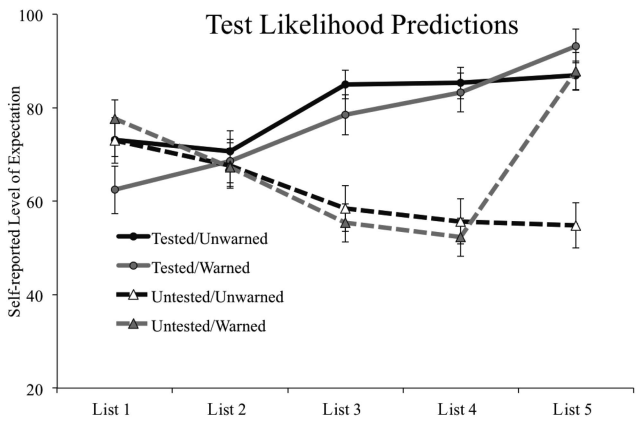

Figure 5. Expectations of a test on the next studied list by testing and warning conditions in Experiment 2 (rating groups). Error bars represent standard error of the mean.

interactions between list number and testing group, $F(3, 486) = 15.2$, partial $\eta^2 = .09$, $p < .001$, between list number and warning group, $F(3, 486) = 13.3$, partial $\eta^2 = .08$, $p < .001$, and among all three variables, $F(3, 486) = 4.25$, partial $\eta^2 = .03$, $p = .006$. The main effect of warning group and the interaction between warning group and testing group were not significant ($ps > .235$). In order to interpret the three-way interaction among list number, testing group, and warning group, we turn to Figure 5. After List 1, there is a clear dissociation between the tested and untested groups such that as list number increases, the tested groups increase in their level of test expectation (although not significantly), whereas the untested groups decrease in theirs. However, this dissociation is violated on List 5 by the untested/warned group, whose test expectation levels mirror those of the tested groups on List 5.

On List 5, tested participants gave higher likelihood ratings than did untested participants, and warned participants reported higher likelihood ratings than did unwarned participants. However, participants who were tested provided ratings approximately equal to those who were untested but warned about the upcoming test on List 5, and only the untested/unwarned group reported lower expectancy ratings than did the other groups. We conducted a 2 (testing group: tested vs. untested) × 2 (warning group: warned vs. unwarned) ANOVA, which revealed a significant main effect of testing group, $F(1, 162) = 20.8$, partial $\eta^2 = .11$, $p < .001$; a significant main effect of warning group, $F(1, 162) = 22.9$, partial $\eta^2 = .12$, $p < .001$; and a significant Testing × Warning Condition interaction, $F(1, 162) = 10.5$, partial $\eta^2 = .061$, $p = .001$.

Initial list tests. Figure 6 shows the number of words correctly recalled on Lists 1–5 for the tested groups and just List 5 for the untested groups. The data are presented as a function of whether participants received a warning prior to study of List 5 but collapsed across test expectancy rating conditions because making ratings of test expectancy prior to the study of each list did not affect correct recall ($ps > 0.485$ for the main effect and all interactions with this variable). Of primary interest was performance on the List 5 test. The correct recall results for the tested versus untested groups were consistent with those found in Experiment 1 and reported in previous articles, with participants performing better when they had been previously tested. As in Experiment 1, we also found that a warning prior to study of List 5 improved performance in the untested condition but not in the tested condition.

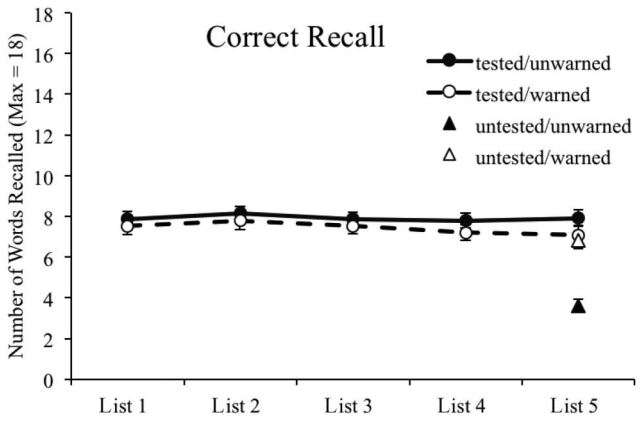

Figure 6. The number of words correctly recalled on each studied list in the two tested groups and on the fifth list in the two untested groups in Experiment 2. Error bars represent standard error of the mean.

The List 5 correct recall analysis was collapsed into a 2 (testing group: tested vs. untested) × 2 (warning group: warned vs. unwarned) ANOVA. This analysis showed a significant main effect of testing group, $F(1, 321) = 30.5$, partial $\eta^2 = .09$, $p < .001$; a significant main effect of warning group, $F(1, 321) = 8.47$, partial $\eta^2 = .03$, $p = .004$; and a significant Testing Group × Warning Group interaction, $F(1, 321) = 24.6$, partial $\eta^2 = .07$, $p < .001$. As in Experiment 1, this interaction indicated that when participants had not received tests on Lists 1–4, a warning prior to study of the fifth list almost doubled performance on that test, $t(159) = 5.89$, $p < .001$ (corrected for unequal variances). On the other hand, when participants had been consistently tested on Lists 1–4, performance was not improved by the warning. In a planned critical comparison, contrary to Experiment 1, the untested/warned group did not perform significantly differently than the tested/unwarned group in terms of correct List 5 recall, although there was numerically better performance in the tested/unwarned group, $t(171) = 1.82$, $p = .07$.

Figures 7a and 7b present data for prior list intrusions for List 5; the two panels differ in terms of whether participants made test expectancy ratings prior to study of each list. Of primary interest was performance on the List 5 test. The prior list intrusion results for the tested versus untested groups were consistent with those found in Experiment 1 and reported in previous articles, with participants producing fewer prior list intrusions when they had been previously tested. As in Experiment 1, we also found that a warning prior to study of List 5 improved performance in the untested condition but not in the tested condition. In a novel finding that differs from correct recall, we also found that making test expectancy ratings prior to study of each list reduced the level of prior list intrusions in the untested groups.

A 2 (testing group: tested vs. untested) × 2 (warning group: warned vs. unwarned) × 2 (rating group: test expectancies rated vs. unrated) analysis showed a main effect of testing group, $F(1, 317) = 107.0$, partial $\eta^2 = .25$, $p < .001$; a main effect of warning group, $F(1, 317) = 32.5$, partial $\eta^2 = .09$, $p < .001$; and a main effect of rating group, $F(1, 317) = 4.65$, partial $\eta^2 = .014$, $p = .032$, on prior list intrusions. Turning first to the testing and warning results, which directly replicate Experiment 1, we found that tested participants produced fewer prior list intrusions, and the warning also reduced prior list intrusions, but this was qualified by an interaction between testing and warning, $F(1, 317) = 26.7$, partial $\eta^2 = .08$, $p < .001$. This interaction indicated that when participants had not received tests on Lists 1–4, a warning prior to study of the fifth list more than halved the number of prior list intrusions on that test, $t(153) = 5.54$, $p < .001$ (corrected for unequal variances). On the other hand, when participants had been consistently tested on Lists 1–4, prior list intrusions were not significantly reduced by the warning. In a planned critical comparison, though, the untested/warned group did produce more prior list intrusions than the tested/unwarned group, replicating Experiment 1, $t(99) = 3.73$, $p < .001$ (corrected for unequal variances).

In addition to the results reported for testing and warning, we also found a significant effect of rating group, such that making test expectancy ratings prior to the study of each list reduced the occurrence of prior list intrusions. Qualifying this conclusion,

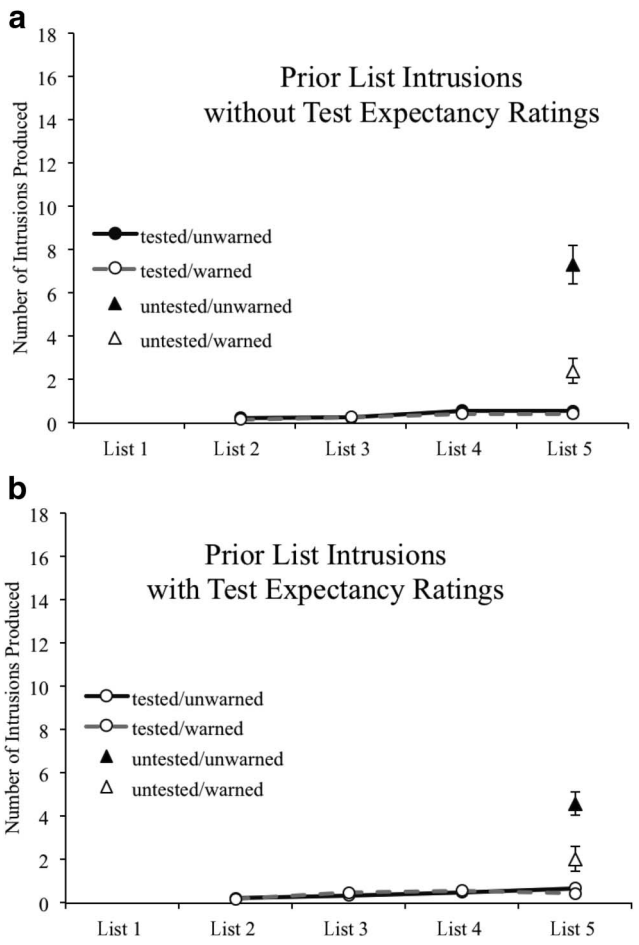

Figure 7. The number of prior lists intrusions produced on Lists 2–5 in the tested groups and on the fifth list in the untested groups in Experiment 2 as a function of whether test expectancy ratings were made. Error bars represent standard error of the mean.

there was also a significant interaction between testing group and rating condition, $F(1, 317) = 5.61$, partial $\eta^2 = .02$, $p = .019$. This interaction reflected the pattern that making test expectancies only helped participants who had not been previously tested avoid prior list intrusions, although it must be noted that participants in the tested/unwarned group already had a very low number of prior list intrusions, so the interaction could be an artifact of floor effects for tested participants (a similar concern arises in relation to the interaction between testing and warning discussed above). The interaction between warning condition and rating condition was not significant ($p = .107$).

Cumulative test. The left side of Figure 8 presents the proportion of words recalled correctly on the cumulative test as a function of testing and warning conditions. Since making test expectancy ratings had no impact on performance on the cumulative test ($ps > .375$ for main effect and all interactions with this variable), we collapsed across this variable for subsequent analyses. For overall performance on the cumulative test, there was only a significant main effect of testing condition, $F(1, 321) = 34.7$; partial $\eta^2 = .10$, $p < .001$, such that, on average, tested groups produced 8.5 more of the 90 words they had studied ($ps > .162$ for the other main effect and interaction). When looking just at recall from List 5 on the cumulative test, as shown on the right side of Figure 8, a different pattern emerged. There were significant main effects of testing, $F(1, 321) = 17.3$; partial $\eta^2 = .05$, $p < .001$, and warning, $F(1, 321) = 4.27$; partial $\eta^2 = .01$, $p = .040$, but these were qualified by an interaction, $F(1, 321) = 12.8$; partial $\eta^2 = .04$, $p < .001$. That is, the untested group recalled more words from List 5 on the cumulative test when they had been warned about the test prior to study of that list, $t(167) = 4.14$, $p < .001$, whereas there was no significant difference between the warned and unwarned conditions among the tested group.

Discussion

In Experiment 2, we directly assessed how groups differed in terms of their expectations of receiving a test after each list. We found that despite all participants receiving the instruction that the presence/absence of a test on each list was randomly determined by the computer, participants in the untested groups showed a clear decrease in their expectation of a test as they advanced through the five word lists. This stands in contrast to the tested groups, whose expectancy from List 1 through List 5 did not significantly change. In addition, we were able to show that introducing a warning prior to study of List 5 shifted test expectancy in the untested group to match that of the tested groups. These data are in line with the assumption we made in our explanation of results from Experiment 1, where we reasoned that participants in the untested group were helped by the warning before studying the fifth list because it brought their test expectancy in line with that of participants in the tested group. An interesting but unexpected finding emerged from the comparison of the two sets of participants—those who made test expectancy ratings prior to each list and those who did

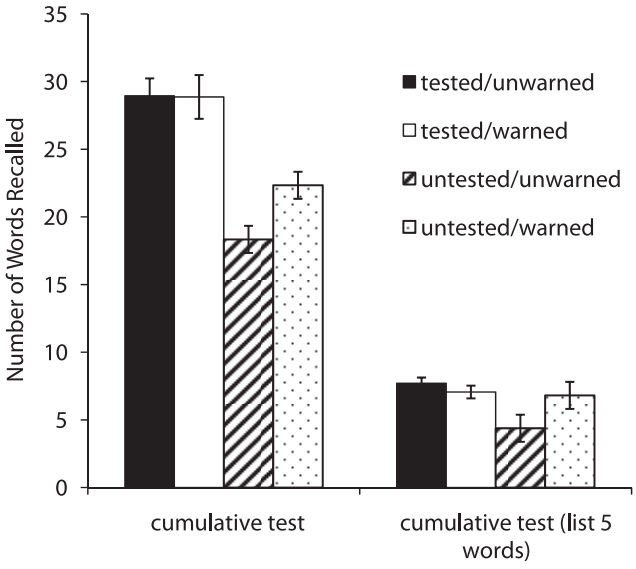

Figure 8. The number of words correctly recalled on the cumulative test in Experiment 2 in the tested × warned groups, from all studied lists (left side), and from the fifth list only (right side). Error bars represent standard error of the mean.

not. Making these ratings appeared to help participants who were not tested after every list to avoid producing prior list intrusions.

One set of participants in Experiment 2—those who did not make test expectancy predictions—performed the same task with the same manipulated variables (testing and warning) as Experiment 1, allowing us to see whether we could replicate the novel Experiment 1 finding that a warning prior to study of List 5 served to reduce proactive interference. Indeed, with a new procedure that excluded participants in all conditions who appeared to disengage from the experiment, we found that a warning about an upcoming test prior to the study of a list enabled participants to produce as many words from that list as those participants who had been consistently tested on prior lists (although prior list intrusions were still less frequent for previously tested participants). In Experiment 2, we once again obtained clear evidence that test expectancy plays an important role in explaining the beneficial effects of interpolated testing, supporting the attentional mechanism explanation that we propose in this article.

General Discussion

In this article, we took two different approaches in an attempt to determine whether the build-up of proactive interference in a multilist paradigm can be attributed, at least partially, to reduced test expectancy. In both experiments, following Szpunar et al. (2008), we included a tested group (these participants received a test after each of five lists) and an untested group (these participants only received a test after the fifth list). In both experiments, we attempted to improve performance by warning half of the participants about the upcoming test before they were given the opportunity to study the last list. In Experiment 1, participants in the untested/warned group performed much better than participants in the standard untested/unwarned group, both in terms of correct recall and prior list intrusions. In particular, correct recall in the untested/warned group was close to that of participants who had been tested on each of the previous lists (whereas the warning did not affect performance in the tested group). In Experiment 2, we excluded participants who could not reproduce the warning about the impending test immediately after reading it. In this experiment, the untested/warned group did not significantly differ from the tested groups in terms of correct recall, although untested/warned participants still performed significantly worse than either tested group in their commission of prior list intrusions. This suggests that testing has at least some benefit above and beyond that provided by expectation, albeit selectively acting to reduce intrusions.

Why might the warning for the untested group have produced a release from proactive interference? Whereas previous accounts of the influence of testing on proactive interference have focused on the role of encoding (Pastötter et al., 2011) and retrieval (Szpunar et al., 2008) mechanisms, we suggest that attentional mechanisms may play a key role in the reduction of proactive interference. Participants in the untested group may encode and retrieve the final list in the study sequence less effectively simply because they are not attending to the task in the same manner as participants who have taken prior tests. Recent work by Szpunar, Khan, and Schacter (2013), in which they reported that intermittent testing reduced mind wandering during online learning, supports the attentional mechanism explanation. The benefit shown by participants in tested groups in previous research (e.g., Szpunar et al. 2008; Weinstein et al., 2011) may in large part be due to the tests serving as repeated reminders to participants that they need to pay attention (i.e., because of the tests they are constantly receiving). It must be noted that our hypothesis does not specify the means by which this test expectancy improves performance; the effect could be related to either attentional processes, quality of encoding, or both. Further, it may be that the tests do not so much encourage people to intentionally decide to attend to subsequent lists as that they enable sustained attention to the lists.

In the paradigm presented here, participants in all groups are told to expect a final cumulative test, which is thought to affect how they approach learning (Szpunar et al., 2007). However, participants who receive a test after every list may also, in the course of the experiment, develop an expectation that a test is likely to follow every list, whereas the untested participants have no such experience to form such an expectation. Our second approach was to test this assumption empirically: In Experiment 2, we obtained test likelihood predictions before every studied list in both the tested and untested groups as a means of directly assessing test expectancy. As predicted, we found that participants in the untested groups gradually lost the belief that they would be tested after studying the next list, while participants in tested groups maintained high levels of expectancy throughout.

By what mechanism might these beliefs affect attention during encoding? Two possible mechanisms, which are not mutually exclusive and, in fact, are likely to co-occur, are list-specific processing and task engagement. When a test is expected after every list, participants may better focus their attention on learning each individual list. Conversely, when learning occurs in the absence of imminent test expectancy, participants may not be performing active, list-specific processing. Evidence that list-relational processing is encouraged by an expectation of a final cumulative test has been demonstrated by Szpunar et al. (2007). In that study, Szpunar et al. showed that, relative to when participants did not expect a final cumulative test, measures of list organization were higher (i.e., participants were more likely to relate information between lists) when participants expected a final test. The expectation of an imminent test (in the tested condition, or after a warning) may encourage some level of list-specific processing over and above the list-relational processing that an expectation of the final cumulative test produces. In support of this hypothesis, we found in both experiments that, relative to untested/unwarned participants, untested/warned participants were better able to not only recall words from the fifth study list but also to avoid generating prior list intrusions.

In addition, participants who are not tested after every list, and thus lose the belief that they will be tested after the next list, may be less engaged in the task. The warning prior to study of the fifth list then serves to re-orient them to their task goals and boosts performance to the level observed in the tested group, who had consistently been tested and expected the test to come. This re-orientation to task goals may also engender the list-specific process described previously. A similar explanation can help account for the unanticipated finding that test expectancy ratings themselves produced a small but significant improvement in terms of reduction of prior list intrusions on List 5 for previously untested participants. Requiring participants to make test expectancy ratings prior to the study of each list may have served as a reminder that

each list could be tested, thus encouraging participants to pay attention. Since this effect was only found for prior list intrusions, this suggests that correct recall and prior list intrusions may be partially subserved by different mechanisms, which would need to be clarified in further work.

A novel aspect of the current article, with regards to this particular paradigm, is that all of our participants were drawn from an anonymous pool and tested online. Although this method of data collection is becoming increasingly popular (e.g., Behrend, Sharek, Meade, & Wiebe, 2011), it is worth noting that in Experiment 2 we excluded 31 participants (roughly 9% of eligible participants) for failing various catch questions throughout the experiment that assessed their processing of experimental instructions. This highlights the importance of manipulation checks when instructions are a key aspect of the design in online samples where we do not have the same level of experimental control as we do in laboratory settings (see Crump et al., 2013; Goodman et al., 2013; and Oppenheimer, Meyvis, & Davidenko, 2009).

The finding that a warning before study of the fifth list can almost bring performance in line with that of participants who had been tested on every list is consistent with recent data by Finn, McDermott, Szpunar, and Wilkie (2013). In their experiments, Finn et al. varied the number of lists participants were given to study prior to taking a test on the most recently studied list. With this design, the authors found that participants always performed poorly on the first test they took after studying multiple lists without interpolated testing, but greatly improved on the subsequent test, performing on that test almost as well (or just as well, in some experiments) as participants who had been tested after every list they studied. One interpretation of these results is that the first test participants received served as a warning that subsequent lists would also be tested, and participants therefore engaged attention during encoding of the following list(s). Of course, it is not possible to determine whether that first test improved performance on later tests via attentional mechanisms or whether testing cleared the proactive interference via some other mechanism. However, Finn et al.'s data are consistent with our finding that a warning about an upcoming test prior to study is more or less equally effective compared with receiving a test after every list, as well as our finding that test expectancy declines when participants do not receive a test after each list. The data we reported strongly suggest that test expectancy and, presumably, attention more generally, play a significant role in the benefit of testing in insulating against proactive interference.

References

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 576–587. doi:10.1037/0278-7393.6.5.576

Bäuml, K. H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language, 68,* 39–53. doi:10.1016/j.jml.2012.07.006

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods,43,* 800–813. doi:10.3758/s13428-011-0081-0

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 431–437. doi:10.1037/0278-7393.33.2.431

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS One, 8,* e57410. doi:10.1371/journal.pone.0057410

Finn, B., McDermott, K. B., Szpunar, K. K., & Wilkie, L. (2013). *Release from proactive interference in long-term memory: The role of retrieval*. Manuscript in preparation.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6,* 1–104.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26,* 213–224. doi:10.1002/bdm.1753

Inda, M. C., Muravieva, E. V., & Alberini, C. M. (2011). Memory retrieval and the passage of time: From reconsolidation and strengthening to extinction. *Journal of Neuroscience, 31,* 1635–1643. doi:10.1523/JNEUROSCI.4736-10.2011

Jensen, O. (2006). Maintenance of multiple working memory items by temporal segmentation. *Neuroscience, 139,* 237–249. doi:10.1016/j.neuroscience.2005.06.004

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 704–719. doi:10.1037/0278-7393.33.4.704

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. doi:10.1016/j.jesp.2009.03.009

Palva, S., & Palva, J. M. (2007). New vistas for α-frequency band oscillations. *Trends in Neurosciences, 30,* 150–158. doi:10.1016/j.tins.2007.02.001

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 287–297. doi:10.1037/a0021801

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General, 106,* 376–403. doi:10.1037/0096-3445.106.4.376

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, USA, 110,* 6313–6317. doi:10.1073/pnas.1221764110

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition, 35,* 1007–1013. doi:10.3758/BF03193473

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1392–1399. doi:10.1037/a0013082

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review, 18,* 518–523. doi:10.3758/s13423-011-0085-x

Received July 28, 2012
Revision received September 25, 2013
Accepted October 4, 2013 ■