Testing the limits of testing effects using completion tests
Scott R. Hinze*, Jennifer Wiley*
* University of Illinois at Chicago, Chicago, IL, USA

Online publication date: 14 April 2011
Testing the limits of testing effects using completion tests

Scott R. Hinze and Jennifer Wiley

University of Illinois at Chicago, Chicago, IL, USA

Recent work on testing effects has shown that retrieval practice can facilitate memory even for complex prose materials (Roediger & Karpicke, 2006a, 2006b). In three experiments the current study explores the effectiveness of retrieval practice on fill-in-the-blank (FITB) tests requiring the recall of specific words or phrases from a text. Final tests included both repeated items that were directly taken from initial tests, and related items. In Experiment 1, with a 2-day delay between initial and final tests, FITB testing benefited performance only on repeated items. In Experiment 2 a 7-day delay between testing sessions led to more robust effects on repeated items. However, once again no benefits were seen for related items. In Experiment 3 the scope of retrieval was varied by comparing FITB tests to paragraph recall tests requiring retrieval of all sentences following a topic sentence. Only the more open-ended recall practice demonstrated improvements in transfer to novel questions. The results suggest that scope or type of processing required during retrieval practice is likely a critical factor in whether testing will have specific or robust benefits.

Keywords: Testing effects; Retrieval processes; Transfer.

A recent flurry of publications has emphasised the potential benefits of frequent testing in educational situations (McDaniel, Roediger, & McDermott, 2006; Roediger & Karpicke, 2006a). It is uncontroversial to argue that testing in classroom contexts can influence learning outcomes as tests can be used to provide summative and formative assessment of student progress, as well as to communicate information about the goals and standards of a course (Black & Wiliam, 1998; Crooks, 1988). Other beneficial effects of testing include when feedback following testing leads to superior learning (Butler & Roediger, 2008; Kang, McDermott, & Roediger, 2007), or when initial testing facilitates future learning (see Szpunar, McDermott, & Roediger, 2008). However, some researchers have made a more specific claim about the direct effects of testing on the learning process: that retrieving information from memory during an initial test directly affects its representation (Roediger & Karpicke, 2006a). As such, testing is proposed to be a more beneficial learning activity than other alternatives including re-studying (Roediger & Karpicke, 2006a). The direct effect of testing on memory has been attributed to the differential processing that occurs during retrieval as compared to re-study, which increases the accessibility of information and makes it less susceptible to forgetting (Karpicke & Roediger, 2007).

Empirical studies have generally supported the claim that post-acquisition testing can lead to memory benefits on a follow-up test. Testing effects are prominent in paired associates learning
(Carpenter & DeLosh, 2005; Carpenter, Pashler, & Vul, 2006) and list-learning tasks (Karpicke & Roediger, 2007; Wheeler, Ewers, & Buonanno, 2003). But they have also been found when information from simulated lectures (Butler & Roediger, 2007) and prose (Kang et al., 2007; Roediger & Karpicke, 2006b) are tested, which is the focus of this investigation.

However, the evidence suggests that testing effects may also depend on the type of initial test that is given. For instance, in word list learning, Carpenter and DeLosh (2006, Experiment 1) demonstrated that free recall led to more robust benefits than recognition, regardless of the final test format. Similar results have been found in studies using more complex materials, with recall tests yielding more robust benefits than recognition tests (Butler & Roediger, 2007; Glover, 1989; Kang et al., 2007). One interpretation of these different sets of findings is, essentially, that recall tasks require more retrieval, which is better for enhancing long-term memory, and that recognition tasks are less effective because they require less retrieval processing.

In this literature researchers have most often employed multiple choice and short answer recall formats for initial tests (e.g., Butler & Roediger, 2007; Chan, McDermott, & Roediger, 2006; Duchastel & Nungester, 1982; Kang et al., 2007; Nungester & Duchastel, 1982). Less is known about other types of open-ended testing that also require the retrieval of information from memory. One candidate technique for open-ended testing that has received less attention involves the completion of sentences in fill-in-the blank (FITB) tests, where specific content from the text is selected and critical words or phrases are replaced by blanks. Although FITB tests have received relatively little attention in laboratory studies (cf. LaPorte & Voss, 1975; McDaniel, Anderson, Derbish, & Morrissette, 2007), FITB tests are a popular activity often employed in end-of-chapter review exercises created by textbook publishers. Critically, this test format still requires the retrieval of information, as participants are asked to fill in the blanks using their memory of the texts they read. Thus, if processes involved in retrieval from memory are the mechanism that underlies the direct benefits of testing, then FITB tests should be more effective than re-exposure conditions that do not require retrieval, at least for specifically tested materials. The first two experiments in this paper test this hypothesis.

Other techniques for open-ended testing include free recall tasks, where participants read a text and then are asked to recall the entire text from memory (see Roediger & Karpicke, 2006b) or recall tasks that provide readers with titles or topic sentences as a prompt, and ask them to generate the remaining sentences. Each of these techniques also requires retrieval from memory, which, according to the retrieval practice account, should lead to improvement over re-exposure conditions. There are some clear differences between these recall tasks and the FITB tests used in our first two studies. The FITB tests are more targeted towards critical material, whereas free recall or paragraph recall require longer, more elaborate responses. While both FITB tests and more open-ended Recall tasks require retrieval, they vary in scope. The goal of the third study is to directly compare the effectiveness of these different kinds of recall tasks, to determine whether the retrieval of information from a memory representation is sufficient for testing effects to be obtained, or whether the scope of the retrieval matters.

A final important issue that is addressed in all three experiments is whether testing information leads to superior performance only on items that were initially tested, or whether the benefits of retrieval practice can be seen to lead to superior performance in new contexts or on materials that were not initially tested (transfer). Anderson and Biddle (1975) argued that if the effects of questioning are limited to repeating the same items verbatim on the final tests, then the effect may be “trivially specific”. Testing effect studies generally test exactly the same material on the final test, often using untested material as the control condition, leaving the question open as to the utility of this mnemonic enhancement for novel tests. More recently some studies have been able to demonstrate improvements beyond just those repeated test items (Butler, 2010; Chan, 2009; Rohrer, Taylor & Sholar, 2010).

In the present experiments we used repeated test items on initial and final tests in order to replicate the traditional testing effect. In addition we explored two categories of novel transfer items. The first category included untested related items in the same test format (cf. Chan, 2009; Chan et al., 2006). The second category included new transfer questions that queried information related to the tested materials, but utilised a different format (cf. Kang et al., 2007; Nungester & Duchastel, 1982, Roediger &...
Karpicke, 2006a). These transfer items were designed to assess the specificity of the benefits of retrieval practice and whether any advantage for initial testing could be translated into the ability to execute that knowledge in a new context.

**EXPERIMENT 1**

After participants had read six science texts, we presented sentences from the texts and asked participants to either fill in blanks with a specific detail from the text or to re-read the sentences. To control for re-exposure (Carrier & Pashler, 1992; Kuo & Hirschman, 1996), participants in the comparison condition read the same sentences again without blanks. In the first experiment participants returned 2 days later, and attempted to fill in both old and new blanks for every text.

The main question addressed by this experiment was: Will FITB retrieval practice yield superior performance on both identical and transfer test items?

**Method**

**Participants.** A total of 79 introductory psychology students from a large urban Midwestern university were recruited for course credit. Of these, 10 participants were omitted because their behaviour and performance (0% accuracy on initial tests in any one condition) indicated non-compliance with task demands. This left a sample size of 69.

**Materials and procedure.** Six expository texts were presented on the following topics: photosynthesis, viruses, glaciers, cell division, corked bats, and acquired heart disease. Text length varied from 391 to 491 words with a mean length of 449 words. After instructions to read for comprehension, all six texts were presented sequentially. Participants read the texts at their own pace.

After all six texts had been read, selected target regions were presented a second time. A target region consisted of a single sentence or pair of sentences that were pulled directly from the text and presented independently. In the re-statement control condition the target region was presented without blanks and the participant was instructed to type the word “read” to confirm that they read the sentences during re-presentation.

In the FITB testing condition two words or phrases of similar length in each target region were identified to be replaced with blanks. The Appendix shows the Cell Division text with example target items. The blanks on the initial FITB tests asked for specific information from the region such as names of components (daughter cells, metaphase plate) and could appear throughout the sentences (i.e., not just at the end of the sentence). “New” blanks on the final FITB test asked for related concepts within the same target area that were relevant to the process of cell division (new cells are “genetically identical”; the cell is “split in two” at the location where the chromosomes align). For the FITB tests participants were instructed to type in the missing word or words even if they had to guess.

The order of text presentation was held consistent across participants, but the order of condition was counterbalanced between participants.

After completing a demographics questionnaire, participants were dismissed and asked to return 2 days later for the second session. They were not told the purpose of the second session. During this second session all of the target regions were presented with both old and new blanks, and participants were asked to type in the missing information from each blank even if they had to guess.

**Design.** Test Item Type was manipulated within participants and had two levels: Repeated Test, where the same information required for the final test had been queried on the initial test (i.e., the information in the “old” blanks); and Related Test, where new information was tested on the final test (i.e., the “new” blanks). In addition there were two levels of initial testing also manipulated within participants: FITB tests and Restatement. With this design we could independently analyse the influence of testing on later memory for that same information by comparing Restatement and FITB testing conditions on repeated test items. This would represent a traditional testing effect. We could also explore whether FITB testing effects transferred to related test items that were read in the same target area.

**Results**

**Initial FITB test accuracy.** Participants generally did not perform at high levels on the initial FITB test. Mean accuracy was \( M = 43.6\% \),
$SD = 21.3\%$. Initial accuracy level is included as a solid horizontal line in Figure 1 to illustrate the estimated amount of forgetting from initial test in comparison to final test performance.

**Final FITB test accuracy.** A 2 (Testing Condition: FITB Test, Restatement) × 2 (Test Item Type: Related, Repeated) repeated-measures ANOVA was performed based on the data shown in Figure 1. Overall, repeated and related test items did not differ in difficulty as indicated by a non-significant main effect of Test Item Type, $F(1, 68) = 0.60, MSE = 0.02$. There was a significant effect of Testing Condition, $F(1, 68) = 7.33$, $MSE = 0.033$, $p = .01$, $\eta^2_{partial} = .10$. As can be seen in Figure 1, the FITB Test condition scored higher on final tests overall. The first critical comparison beyond this main effect is between performance on Repeated test items in the FITB Test condition and performance on the same items in the Restatement control condition, which provides the most direct check for a testing effect. Indeed we were able to replicate the typical testing effect here, $t(68) = 2.88$, $p = .01$, $d = 0.42$. Next, we checked for transfer to related items by comparing performance on Related items in the FITB Test condition with performance on the same items in the Restatement control condition. This comparison was in the expected direction, but was not significant, $t(68) = 1.18$, $p = .24$, $d = 0.15$. While the pattern suggests that the effect of testing differed for repeated and related items, a test of this interaction was not significant, $F(1, 68) = 1.99$, $MSE = 0.024$, $p = .16$, $\eta^2_{partial} = .03$.

It is also important to note the level of initial test performance shown in Figure 1 by the solid horizontal line. The difference between this line and final test performance serves as an estimate of forgetting between the initial test and final test. It appears that there was substantial forgetting in the Restatement condition (but not the FITB condition) for repeated items, which contributed to the direct testing effect. However, there was little forgetting of the related test items in either group, so it may not be surprising that there was no difference between testing conditions on the related items.

**Conditional probabilities.** In order to further explore the influence of FITB testing on retrieval during final tests, we analysed initial test accuracies for the testing conditions. It is well established that initial tests are more effective for items that are successfully recalled during initial tests than for those not successfully recalled (e.g., McDaniel & Masson, 1985; Thompson, Wenger, & Bartling, 1978). This trend may be especially important for the current experiment due to the relatively low levels of initial test performance (see Kang et al., 2007). Conditional analyses may demonstrate whether higher initial accuracy would lead to higher final test performance for either repeated or related content.

Final test performance was analysed using a 2 × 2 repeated-measures ANOVA with initial retrieval success (successful, unsuccessful) and final test content (repeated, related) as within-participants variables. As would be expected there was a main effect of initial retrieval success.
with successful retrievals leading to higher final test performance, $F(1, 65) = 140.81$, $p < .001$, $\eta^2_{\text{partial}} = .68$. However, this main effect was qualified by a strong interaction with final test content, $F(1, 65) = 134.18$, $p < .001$, $\eta^2_{\text{partial}} = .67$. Follow-up comparisons demonstrated that there was a substantial benefit for successfully retrieved repeated items ($M = 0.84, SD = 0.23$) over unsuccessful retrievals of repeated items ($M = 0.14, SD = 0.18$), $t(67) = 18.29$, $p < .001$. In contrast, successful initial retrievals were associated with only marginally higher performance for related items ($M = 0.50, SD = 0.34$) as compared to unsuccessful initial retrievals ($M = 0.42, SD = 0.30$), $t(67) = 1.66$, $p = .10$. This interaction suggests that the benefits of successful retrieval are very strong for content that is directly tested, but that being able to retrieve one part of the question prompt only had a marginal influence on memory for the related content, at least after a 2-day delay.

**Discussion**

This experiment suggests that the effects of FITB retrieval practice may be substantial but very specific. Testing effects were only found on repeated test items, with no significant facilitation of related information when compared to restatements.

**EXPERIMENT 2**

After a 2-day delay, tested information was better remembered than re-studied information (a positive testing effect), but no transfer effects were seen on related information that was presented in the same target area. In a second experiment we explored whether a longer delay between initial and final tests would allow for transfer effects to emerge.

Some have suggested that retrieval practice serves to attenuate the forgetting of information (Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006b; Thompson et al., 1978; Wheeler et al., 2003). With only a 2-day delay there might have been relatively little forgetting in the restatement condition, which could have left no opportunity for retrieval practice to affect the relative accessibility of information. Following the logic that testing effects may be obscured with shorter delays, increasing the delay between initial and final tests should allow for more forgetting. Once substantial forgetting is obtained, then retrieval practice may lead to greater accessibility of relevant information for both repeated and related items, and initial testing should outperform restudy.

**Method**

A total of 30 introductory psychology students participated in this experiment. Of these, 4 participants were omitted due to non-compliance on initial tests, leaving a sample size of 26. The experimental design and all materials were identical to Experiment 1. The procedures were also identical, except that, after being dismissed from the initial session, participants returned after 7 days, rather than after 2 days.

**Results**

**Initial FITB test accuracy.** Initial accuracy was similar to Experiment 1. Mean accuracy was $M = 44.7\%$, $SD = 22.4\%$.

**Final FITB test accuracy.** As in Experiment 1, a 2 (Test Condition: FITB Testing, Restatement) $\times$ 2 (Test Item Type: Related, Repeated) repeated measures ANOVA was performed based on the data shown in Figure 2. Overall, repeated and related items did not differ in difficulty, as there was no main effect of Test Item Type, $F(1, 25) = 0.03$, $MSE = 0.02$, ns. There was a significant main effect of Testing Condition, $F(1, 25) = 6.52$, $MSE = 0.011$, $p = .02$, $\eta^2_{\text{partial}} = .21$. As can be seen in Figure 2, the FITB Testing condition outperformed Restatement overall. In this experiment the effect of Testing Condition was moderated by a significant interaction with Test Item Type, $F(1, 25) = 5.90$, $MSE = 0.02$, $p = .02$, $\eta^2_{\text{partial}} = .19$. The first critical comparison to follow up this interaction was between performance on Related items in the FITB Test condition and performance on the same items in the Restatement control condition, which provides the most direct check for a testing effect. As in Experiment 1 we were able to replicate the typical testing effect, $t(25) = 4.03$, $p < .001$, $d = 0.69$. The second critical comparison, between performance on Related items in the FITB Test condition and performance on the same items in the Restatement control condition, was not significant, $t(25) = -0.37$, $p = .71$. 

Overall, it appeared that increasing the delay between initial and final tests did increase forgetting. Again, a testing effect was found on repeated test items and the effect sizes suggest the effect was more substantial with the longer delay ($d = 0.42$ after 2 days, $d = 0.69$ after 7 days). However, as with Experiment 1, we failed to find significant effects of retrieval practice for performance on related items, as confirmed by the significant interaction in Experiment 2. That is, the longer delay led to more forgetting for the related items regardless of testing condition.

**Conditional probabilities.** We also calculated conditional probabilities as in Experiment 1 to determine the levels of performance on final tests based on successful initial retrieval. In Experiment 1 successful initial retrieval had a marginal benefit for related information, but this may have been attenuated by the lack of forgetting of the related materials. Presumably, if initial retrievals have an influence on related content and substantial forgetting has occurred for the related content in general, successful retrievals should lead to more facilitation of related content than unsuccessful retrievals. We tested this with a $2 \times 2$ repeated-measures ANOVA with initial retrieval success and final test content as withinparticipants variables.

As in Experiment 1, there was a strong effect of retrieval success, $F(1, 22) = 36.01$, $p < .001$, $\eta^2_{partial} = .62$, and a strong interaction between success and final test content, $F(1, 22) = 42.00$, $p < .001$, $\eta^2_{partial} = .66$. Successful recall led to higher final test performance for repeated content ($M = 0.73$, $SD = 0.29$) as compared to unsuccessful retrieval ($M = 0.10$, $SD = 0.14$), $t(25) = 9.51$, $p < .001$. In contrast, related test performance was equivalent after successful retrieval ($M = 0.36$, $SD = 0.32$) as compared to unsuccessful retrieval ($M = 0.34$, $SD = 0.30$, $t < 1$). As with Experiment 1, related content did not seem to be substantially influenced by initial retrieval attempts, even when those attempts were successful.

**Discussion**

In two consecutive experiments we demonstrated a robust effect of FITB retrieval practice during initial testing events on memory for information acquired from short expository science texts. While some other experiments have found an advantage for initial FITB testing only with feedback (Kang et al., 2007; LaPorte & Voss, 1975; McDaniel et al., 2007), the present direct testing effects were found even without feedback. It is likely that these effects could be even larger if the indirect effects of feedback were added. Importantly, these findings were also demonstrated in contrast to comparison conditions that were matched for re-exposure. Previous studies that have demonstrated direct effects of completion testing have used control conditions that did not allow participants to re-study information (e.g., Glover, 1989). This means that some of the facilitation could have been due to re-exposure to textual information from the question stems in the testing condition. Thus the results of the current studies add to the literature suggesting...
that FITB retrieval practice can aid later memory on final tests involving the exact same items, even without feedback and with a comparison condition that controls for re-exposure.

While the direct effects of testing were robust and encouraging, the benefits of initial FITB testing on final test performance were only seen when the exact same blanks were tested during both tests. No evidence was seen that FITB retrieval practice facilitated performance on the related test items. These related items required memory for information from the same target area that was not queried on initial tests. This information was clearly related not only to the general topic, but also to the specific information that was queried on initial tests. However, no effects of retrieval practice were found on these items.

The failure to find improvements on the related items may be a function of the re-exposure control conditions that were used in these experiments. Previous studies that have demonstrated “retrieval-induced facilitation” for related test items (Chan, 2009; Chan et al., 2006) have used control conditions that did not allow participants to re-study the related information. While those previous studies found benefits of the activation of related information due to testing, the current experiments allowed participants in both the restudy and testing conditions to re-read the related information. Perhaps due to this design feature, no benefits of testing were found in relation to this re-exposure condition.

Another possible reason for the failure to find facilitation on related items on final tests could be the low initial test accuracy found in both studies. That is, it is possible that if participants had been more successful at initial retrieval, benefits for related items on final tests may have emerged. While this is an empirical question, the conditional analyses demonstrated that relative success on initial tests had little effect on related items. This casts some doubt on the hypothesis that initial test performance caused the failure of testing effects to transfer to related items.

A third explanation for the observed pattern of results is that the FITB items used here may only require surface-level memory for text, or may only require processing of information immediately surrounding the blank in order to fill in an appropriate answer. Thus this FITB format may have particularly prompted superficial, item-specific processing. This interpretation of the observed results would suggest that the nature of the processing required during different types of initial tests might be critical to explore. Experiment 3 explores this issue by analysing the robustness of FITB testing effects and by comparing FITB tests with more open-ended recall tasks.

**EXPERIMENT 3**

The first two experiments, while firmly establishing the presence of direct effects of FITB testing on repeated items, suggest the possibility of a “trivially specific” (Anderson & Biddle, 1975) effect with little evidence of transfer as compared to re-exposure. Even though FITB tests require retrieval, this was not sufficient to lead to improved performance on related items. Experiment 3 explores the possibility that the scope or type of retrieval practice required by an initial test may matter. In addition, Experiment 3 focuses on a different type of transfer; that is, whether the benefits of testing hold for initially tested content queried with novel test question formats (Kang et al., 2007; Roediger & Karpicke, 2006a).

To test the impact of retrieval scope, we manipulated the extensiveness of the generation required by the initial tests. In one condition, similar to Experiments 1 and 2, FITB tests required the generation of single words or short phrases. Based on free recall designs (e.g., Roediger & Karpicke, 2006b), a second condition involved recall of full portions of the text as the initial test. In this condition participants were cued with the first sentence of each paragraph and engaged in free recall on the remainder of each of those paragraphs. In a third condition participants engaged in the same recall task with the opportunity to restudy the text after recall. This sort of recall with restudy has been shown to transfer to new questions in at least one experiment (McDaniel, Howard, & Einstein, 2009; Experiment 2), so we expected this condition would have the most opportunity for success. In addition, two comparison conditions were included; one in which participants simply read the texts once with no testing or opportunity for re-reading, and one that allowed the participants to re-read each text a second time. This allowed for the additional assessment of the effects of re-reading.

A prediction based in the retrieval practice account might suggest that engaging in both FITB tasks and in paragraph recall should strengthen
the accessibility of ideas from the target text, which should facilitate final test performance. That is, if availability of specific information in memory is a prerequisite for transfer, then any task that enhances the retrievability of this relevant information should lead to superior transfer (see Butler, 2010).

On the other hand, there are clearly differences in the breadth or depth of retrieval that is required to complete FITB tasks and paragraph recall tasks, and this could influence the kind of testing effects that may be obtained. One way of conceptualising this is through the “desirable difficulties” framework (Bjork, 1994), which suggests that more effortful successful processing should lead to greater learning. In testing effects not only is retrieval a more effortful process than re-exposure, but also more difficult (but successful) retrieval tasks are more effective for long-term retention than less difficult successful retrieval tasks (Carpenter, 2009; Carpenter & DeLosh, 2006; Pyc & Rawson, 2009). This would suggest that paragraph recall tasks, because they require more extensive generation, could lead to more robust testing effects than FITB tasks.

Similarly, another difference between FITB tasks and paragraph recall tasks is that they might require different types of retrieval processes. Previous results suggest that improved performance on later tests might be moderated by the type of processing at initial test (Chan et al., 2006; Johnson & Mayer, 2009; McDaniel & Masson, 1985). In the case of this experiment it may be that FITB tests do not require processing beyond the surface form of the text to retrieve the missing information. Thus FITB tests would not facilitate transfer to new testing formats. On the other hand, to the extent that recall tasks facilitate elaboration and integration of sentences, rather than simply the retrieval of individual words (Kintsch, 1998), then an initial recall attempt involving the generation of multiple sentences might facilitate later transfer performance.

**Method**

**Participants.** A total of 26 Introductory Psychology students completed all aspects of the task for course credit. One participant’s data were removed because of non-compliance with task demands. The remaining 25 participants were evenly represented in the five task orders described below.

**Texts.** Five short explanatory science texts were used as target materials. These texts were adapted from a middle school science textbook (Coolidge-Stoltz et al., 2001). The topics were: endocrine system, vision, respiratory system, viruses, and the water cycle. The texts ranged in length from 335 to 432 words and were presented as a whole. The new passages were used because each had clear topic sentences to serve as retrieval cues and each had five questions designed to serve as final test content.

**Initial study/test conditions.** After reading each text, participants engaged in one of the five initial study or testing exercises listed here (one for each text). The influence of these exercises on later performance was the focus of this experiment.

1. **Read Once.** Participants were not required to re-study or retrieve information from the text, and moved directly to reading the next text.
2. **Re-read.** Participants read the entire text again with the instruction, “Please read the text again. Once again, you should attempt to comprehend the text as a whole.” This condition served as a re-exposure control, since part of the advantage of retrieval may be due to the additional re-exposure of materials (Carrier & Pashler, 1992; Kuo & Hirschman, 1996).
3. **Fill-in-the-Blank Test (FITB).** Participants saw the entire text again with selected sections of text missing and replaced by an underlined blank. They were asked to “Fill in the blanks below as best as you can. You can use your own words if you do not remember the exact words.” Blanks were chosen to represent missing details. Each final transfer item (discussed below) was represented by one of these detailed blanks.
4. **Paragraph Recall.** Topic sentences for each major paragraph were presented with a space for open-ended responses and participants were instructed, “Below you will see topic sentences from many of the paragraphs in the previous text. Please attempt to recall the content from those paragraphs. You may use your own words.”
5. **Paragraph Recall with Re-read.** This condition was identical to paragraph recall, except that, after recall, participants had an opportunity to re-read the text and were instructed, “Now that you have attempted to
recall, please read the text again. Pay attention to the materials that you may have forgotten or left out of your recall.” This condition served as a test of the influence of restudy opportunities in addition to retrieval practice.

Final transfer tests. The final transfer tests consisted of novel multiple choice questions. Five transfer questions were created for each text. Transfer questions tested memory for information directly stated in the text using a new test format. For example, the Viruses text stated, “the cells that viruses infect in order to multiply are called host cells.” In the initial FITB prompt the words “host cells” were left blank. The transfer item for this blank was:

The cells that a virus infects are called what?

A. Support Cells
B. Lock Cells
C. Host Cells
D. Protein Cells

For all questions, there was only one correct answer, and none included options for “all of the above” or “none of the above”.

Procedure. All aspects of the task were completed using paper-and-pencil presentation in small groups. Participants were initially given the following instructions:

In this study you will read five short texts about science topics. After reading some texts you will perform a short activity. Sometimes this activity will require you to recall words or sections of the text. Other times you will re-read the text. You will not know which activity you will complete ahead of time. For each of these you will not be allowed to revisit the original text, so you should read each text carefully the first time.

After these instructions participants alternated reading and initial study/test exercises. The order of texts was always held constant (endocrine system, vision, respiratory system, viruses, and water), but the study/test exercises were counterbalanced so each text was matched with each exercise an equal number of times. Texts were presented as a whole, with instructions for the exercise on a separate page. All exercises occurred immediately after initial reading, so that participants read a text and immediately engaged in the appropriate task before reading the next text.

After completion of the final exercise students were dismissed and were reminded to return to the same location after a 2-day delay. They were not informed of the nature of the second session. At the beginning of session II, participants received test booklets consisting of multiple-choice questions for all five texts. Test topics were presented in the same order as the texts had been presented.

Results

A repeated-measures ANOVA was performed with Initial Study/Test condition as the independent variable and Final Test Performance as the dependent variable. The relevant data for this analysis are presented in Figure 3.

A main effect of initial study/test condition was significant, \( F(4, 80) = 3.39, p = .01, \eta^2_{partial} = .15 \). Pairwise \( t \)-tests confirmed that FITB, Re-read and Read Once did not differ from each other (all \( ts < 1 \)). Also, there was no significant difference between Paragraph Recall and Paragraph Recall with Re-read indicating that any benefit from re-exposure to the text after testing was minimal, \( t(24) = .07 \).

Demonstrating a testing effect, Paragraph Recall statistically differed from Read Once, \( t(24) = 2.28, p = .03, d = 0.71 \), and Restudy conditions, \( t(24) = 2.20, p = .04, d = 0.63 \). Paragraph Recall with Re-read also showed a trend towards an advantage over Read Once, \( t(24) = 1.97, p = .06, d = 0.55 \), and Re-read conditions, \( t(24) = 1.75, p = .09, d = 0.45 \). Since Read Once and Re-read did not differ, and Paragraph Recall and Paragraph Recall with Re-read did not differ, it may be reasonable to consider these conditions as two classes: control conditions without the opportunity to retrieve and experimental conditions with a paragraph-level retrieval opportunity. This conceptualisation allows us to average across these conditions and eliminate some error due to text and ordering. The combination of the Paragraph Recall groups had clearly superior final transfer performance (\( M = 0.75, SD = 0.19 \)) to the combination of the control groups (\( M = 0.62, SD = 0.16 \), \( t(24) = 2.51, p = .02, d = 0.74 \). The combination of Paragraph Recall groups was marginally superior to FITB testing, \( t(24) = 1.87, p = .07, d = 0.50 \), but FITB did not
differ from the combination of control conditions, $t < 1$. In sum, while FITB testing did not lead to superior performance on new MC questions relative to control, Paragraph Recall tasks did.

**Initial test accuracy.** In order to further explore the influence of testing on retrieval during final tests, we analysed initial test accuracies for the testing conditions. In this experiment participants each completed initial tests for three readings: a FITB test and two paragraph recall tests with or without a re-reading opportunity. Based on the five final test questions for each of those texts, each participant could have either successfully or unsuccessfully provided the relevant information in the initial tests, creating a within-participants variable of initial recall success. By basing initial retrieval scores for all conditions only on the relevant final test content, FITB and recall conditions can be compared, since both conditions allowed the possibility of retrieving the relevant content. Because there was no difference between recall and recall with re-reading, those two conditions were combined in order to allow for larger data samples.

The first question was whether test formats would yield different levels of initial test success. That is, did participants perform better after paragraph recall tests simply because they were more successful at retrieving test-relevant content during initial tests? In fact, initial test data showed the opposite pattern. Overall, participants successfully retrieved more test-relevant content in the more focused FITB test format ($M = 0.62, SD = 0.29$) than in the more open-ended paragraph recall tests ($M = 0.40, SD = 0.19$), $t(23) = 4.57, p < .001$. This may not be surprising, since FITB tests cued participants to a single, relevant, word or phrase, rather than an entire paragraph. However, it does indicate that paragraph recall conditions did not simply allow participants to retrieve specific information more successfully.

**Conditional probabilities.** The analyses below explore conditional probabilities within test formats. That is, we tested the likelihood of correctly answering final test questions given successful or unsuccessful retrieval during initial tests. Participants were only included in these conditional analyses if they provided both successful and unsuccessful initial retrievals during initial tests.

**FITB tests.** Given the small sample of possible responses (only five items of interest), only 18 participants provided both successful and unsuccessful initial retrievals. Five participants were successful for all initial FITB items and subsequently performed well on final tests ($M = 0.79, SD = 0.12$). Two participants were unsuccessful on all items. One of these unsuccessful participants incorrectly answered all final test questions, while the other correctly answered two out of five final test questions.

Among the remaining participants with both successful and unsuccessful initial retrievals, unsuccessful retrieval was related to lower final test performance ($M = 0.41, SD = 0.48$) than successful retrieval ($M = 0.71, SD = 0.36$), $t(17) = 2.41, p = .03$. The advantage on final tests for retrieved
content suggest that successful FITB testing had some influence on memory for items participants were able to retrieve.

Paragraph recall tests. Out of 25 participants, 24 provided both successful and unsuccessful retrievals during initial paragraph recall tests. Among these participants, unsuccessful retrieval was related to lower final test performance ($M = 0.70, SD = 0.21$) than successful retrieval ($M = 0.86, SD = 0.21$), $t(23) = 2.69, p = .01$. The advantage for retrieved content suggests that retrieving specific information may be important for final test performance. However, the relatively high performance even for information that was not provided in recalls suggests that the benefits of recall tasks were somewhat more general.

Discussion

The main finding from this third experiment was that a paragraph recall test, which required more extensive retrieval than FITB tests, was able to yield a testing effect on novel multiple choice questions. This demonstrates, once again, that positive effects of testing can be observed even without feedback and in comparison to a re-exposure control condition. However, this study also illustrated some limitations of initial FITB tests as learning events. There was some evidence in the conditional analyses that successful FITB retrievals led to better performance on later MC items. This shows a potentially small transfer effect of FITB tests to MC formats. This effect may seem to be in conflict with the conditional analyses from Experiments 1 and 2, but the transfer measure in this experiment was different (i.e., a multiple choice test on the same content, rather than a related piece of information on a cued recall task). However, the overall effect of prior FITB testing was not significantly better than re-exposure. There may be conditions under which a benefit from prior FITB testing could be demonstrated (e.g., by adding feedback or repeated testing). We note, however, that a simple procedural change from FITB to a more open-ended initial test format had a clear impact on the ability to answer new questions on the final MC test. In the general discussion we return the issue of the superiority of more open-ended testing.

One major empirical contribution of the present experiments was determining a robust testing effect for FITB tests on repeated questions even when compared to a re-exposure condition and when participants were not provided with feedback. Interestingly, in previous studies employing conditions without feedback, neither LaPorte and Voss (1975) nor Kang et al. (2007) found a direct improvement for tested material over a restatement control. In this way the current study has revealed a more robust direct effect of FITB testing for prose than has been found previously.

A second consistent finding across the studies was the failure to find any benefit of initial FITB testing on final test performance beyond identical test items. No benefits were seen for related test items or for novel question formats. Although the first two studies had low initial test performance levels on the FITB tests, the third study had more typical levels of test performance and limited effects of FITB tests remained. This observation, along with the conditional analyses, suggests that the failure to find facilitation on novel items on the final test was not simply due to low initial test performance. However, the role of initial test performance as a mediator of testing effects when learning from text deserves to be investigated more directly in future studies.

The failure of initial FITB tests to facilitate performance on later MC tests could be interpreted to suggest that FITB tests may be less likely to transfer than other formats such as SA or MC tests. Some researchers have found that SA questions facilitate later MC questions but not vice-versa, although these experiments have utilised feedback (Kang et al., 2007; McDaniel et al., 2007). Notably, without feedback, Nungester and Duchastel (1982) were able to demonstrate that performance on initial SA test questions facilitated performance on final MC tests, and vice versa, in a study examining memory for historical facts from a text about the Victorian era. However, it is possible that the facilitation across the two test types in this particular study shared the same question stems such as “What nationality was Prince Albert?” At the same time, at least one other study without feedback has failed to find facilitation from initial MC testing on final SA test performance compared to a re-study condition (Butler & Roediger, 2007). Butler and Roediger (2007) did not examine the effects of an initial SA
test on later MC test items, and no other studies on text have reported transfer of testing effects across formats without providing feedback.

The third major empirical finding in this research was the demonstration that an initial recall test led to more robust benefits of prior testing on learning from text than an initial FITB test. This result extends a similar effect that has been demonstrated with word lists (Carpenter & DeLosh, 2006, Exp 1) into a context where the materials to be learned were expository science texts. It also extends the results of Glover (1989), Experiment 4), which suggested that initial attempts to free recall a text can lead to superior final test performance over initial cued recall attempts. The results from the current Experiment 3 are consistent with Glover’s findings but also extend them by comparing them to a re-exposure control condition. Further, the previous Glover study conflated spacing and testing, which complicated their interpretation, and the current Experiment 3 does not have this issue.

We suggest that this overall pattern of results, facilitation from initial recall tests and limited benefits from initial FITB tests, can be explained by the nature of the processing that each kind of initial test may require. Similar to explorations of the generation effect (e.g., deWinstanley, Bjork, & Bjork, 1996), facilitation of learning outcomes may depend on the type of processing engaged in during the initial testing event. From this perspective it is important to consider the kind of processing required by our FITB tests. In particular, the FITB test items used here were verbatim excerpts pulled directly from the text in which participants were asked to retrieve information for blanks that could appear anywhere in a sentence. Even though these test items required the “production” of a response, the response could be produced with surface memory for the text that was read, instead of a broader retrieval attempt that might involve considering the missing concept in the context of the scientific process that was being described. Thus we suggest that the FITB test items used here may have prompted superficial or item-specific processing during the testing event, which limited their impact on related information or novel test formats. It is possible that, in contrast to the FITB test items used here, more open-ended SA type questions used by others (e.g., Chan et al., 2006; Kang et al., 2007) might require more processing of not only the item to be retrieved, but also the relationship between the item and its context in the text. It is difficult to make conclusions about our proposed distinction between FITB and SA questions across experiments, because other experiments (Chan et al., 2006; Kang et al., 2007) have used recall tests that utilised some combination of FITB and short answer items.

Another intriguing possibility suggested by a reviewer is that the FITB tests used here, rather than facilitating the retrieval of untested but related concepts, might have actually resulted in competition for attention between the tested and untested information. The format of the FITB tests might have caused readers to focus their attention very locally and specifically towards the part of the sentence where the blank appeared, which might have detracted attention from further processing of the related information that was actually present on the screen. In the restudy condition, in contrast, the readers might have allocated more equal amounts of attention to both regions. Thus the FITB test format might have led to a relative disadvantage in the amount of processing of related information compared to the restudy condition, which could have masked any facilitative effect of testing.

In contrast with the failure to find transfer effects for initial FITB tests in the current study, Chan et al. (2006) found facilitation from initial SA tests on novel items that were related to the tested materials, even without facilitation of directly tested items when compared to a partial re-exposure control condition. In that study participants either were tested or re-read one half of the items. They were then tested on all the items on the final test. Initial testing led to no advantage on the items that had been tested or read, but it did lead to an advantage on the items that had been not tested. The authors termed this benefit of initial testing on related items “retrieval induced facilitation” (Chan et al., 2006). Importantly, this effect was seen most clearly in one particular condition: when instructions explicitly prompted readers to engage in a broad search of memory during the initial testing event. Those students who were instructed to perform a narrow search of memory while completing initial SA items experienced no retrieval-induced facilitation on the final test. In other words, the “narrow search” results were consistent with what we observed from the FITB tests used in this study. Thus the retrieval-induced facilitation results can be seen as highly consistent with the findings from our Experiment 3 where we
manipulated the scope of retrieval with two different types of initial tests. Requiring the recall of full paragraphs likely necessitates a broader search of memory than does filling in individual blanks. In this way the current results suggest that retrieval induced facilitation may depend on the nature of retrieval processes engaged in during the initial testing event.

While we were encouraged to find benefits of open-ended retrieval practice extending to new test items and believe that such open-ended retrieval is useful for long-term retention, we note that the kinds of transfer questions used here were still quite limited. The “transfer” tests were essentially re-phrasings of propositions from the text presented in a multiple-choice test format. The current experiments demonstrate no evidence that testing helps participants when they are asked to apply this knowledge in a new context or make inferences based on the content learned from the text.

It is possible that different forms of retrieval practice may facilitate superior processing and deeper levels of comprehension (Kintsch, 2005). For instance, free recall of the entire text, to the extent that it encourages elaboration and integration of ideas, may lead to better performance on inferential questions (McDaniel et al., 2009, Experiment 2). However, others have suggested that deeper understanding is unlikely to result from retrieval practice alone (Johnson & Mayer, 2009). Possibilities for using tests to promote deep understanding and transfer should be the focus of future work.

CONCLUSIONS

It is well established that retrieval practice can have strong mnemonic benefits, and the current experiments support the conclusion that testing can improve memory for short scientific texts. The present results suggest that differentiating the scope or type of processing required by different test formats, and testing for transfer of benefits to new items, are critical issues for further study. Further, these results suggest that differences in retrieval processes during the initial test will likely moderate the utility of testing for different learning outcomes.

Manuscript received 5 June 2010
Manuscript accepted 25 January 2011

REFERENCES


**APPENDIX**

**Sample text: Cell division**

Cell division occurs to reproduce and replace cells. The first stage of cell division is called mitosis. In mitosis, new cells, or daughter cells, are created from a parent cell. Each new cell contains a complete set of chromosomes which guarantees that they are genetically identical. Mitosis has four distinct phases. The first is called prophase (pro- means “before”). During prophase, invisible, thread-like DNA fibers of the nucleus, which are called chromatin, condense and double into two chromosomes, each of which is called a chromatid. Each chromatid pair is attached at one place, to form a single chromosome. The place where these chromatids are attached is called the centromere. Soon after the chromatin material has condensed into doubled chromosomes, centrioles begin to migrate away from each other. Centrioles are two small structures located outside the cell's nucleus. Thread-like filaments develop around each centriole and radiate in all directions, resembling a flower. During later prophase, many of the filaments between the two centrioles lengthen and connect with each other. This network of filaments is called the spindle. In human cells, prophase lasts from 30 to 60 seconds.

The second phase of mitosis is called metaphase (meta- means “mid”). During metaphase the chromatids become aligned at the midregion, or equator, of the cell. When the chromatids align this forms the metaphase plate which later becomes the location where the cell is split in two. At this time the centrioles are at opposite
ends of the cell, which are called the poles. Also during metaphase, the formation of the spindle between the two centrioles is completed. Metaphase in human cells takes two to six minutes.

At the beginning of the third phase of mitosis, called anaphase (ana- means “away”), the protein structures located within the centromere of the chromatids all divide at one time. This division causes the chromatids to separate into daughter chromosomes. The spindles then pull the daughter chromosomes to each cell pole, which is why this is called the “away” phase.

The fourth and final stage of mitosis is called telophase (telo- means “end”), and it begins when all the daughter chromosomes reach the two cell poles. During telophase the spindle that was completed in metaphase begins to disappear. Later, a single nuclear membrane reappears and encloses the two groups of chromosomes at the two poles. While this is happening, the chromosomes begin to disappear and turn back into threadlike chromatin material, or DNA, which spreads throughout the nucleus. Telophase in humans is quite variable, requiring from 30 to 60 minutes.

Sample target areas

Tested items from each target area are presented in brackets. During initial testing one item was blank. During final testing both items were blank.

In mitosis, [daughter cells] are created from a parent cell. Each new cell contains a complete set of chromosomes which guarantees that they are [genetically identical].

When the chromatids align this forms the [metaphase plate] which later becomes the location where the cell is [split in two].