Journal of Memory and Language 69 (2013) 151–164

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

The importance of constructive comprehension processes in learning from tests

Scott R. Hinze *, Jennifer Wiley, James W. Pellegrino

Department of Psychology, University of Illinois at Chicago, United States

ARTICLE INFO

Article history:
Received 24 July 2012
revision received 31 December 2012
Available online 20 April 2013

Keywords:
Testing effect
Retrieval processes
Comprehension

ABSTRACT

The goal of these experiments was to introduce and test the constructive retrieval hypothesis, according to which retrieval practice will be most effective when it encourages constructive elaborations of text content. Experiment 1 provided baseline performance data for the materials included in Experiments 2 and 3. In Experiment 2, instilling inference-based test expectancies before an initial retrieval attempt led to more constructive retrieval practice and better final test performance than instilling detail-based expectancies. In Experiment 3, instructions to construct explanations during initial retrieval attempts led to more constructive retrieval practice than free recall, and better final test performance than free recall or rereading instructions. These experiments support a constructive retrieval account of testing effects, and demonstrate that it is not retrieval practice alone, but rather the kind of constructive processing invoked during retrieval attempts that can improve both retention and comprehension when learning from text.

Introduction

The idea that simply testing one's own memory can serve as an effective learning activity is an appealing one and recent research among cognitive psychologists suggests that retrieval practice may indeed be useful (see Roediger & Karpicke, 2006a, for a review). The advantage of testing over a re-study opportunity has been designated the "testing effect" and evidence for this advantage has been cited as a reason to increase the frequency of testing in classrooms (McDaniel, Roediger, & McDermott, 2006; Pashler et al., 2007; Roediger, Agarwal, Kang, & Marsh, 2010).

The term "testing effect" requires clarification, however, as there can be many effects of tests (see Crooks, 1988), from shaping future study (Mawhinney, Bostow, Laws, Blumenfeld, & Hopkins, 1971; Szpunar, McDermott, & Roediger, 2008), to inducing anxiety (Hembree, 1988), to allowing for feedback (Butler & Roediger, 2008) and formative interventions (Black & Wiliam, 1998; Wiliam, 2007). What is at issue in research on the testing effect is not simply whether tests can be useful in learning contexts. Rather, proponents of applying testing effects in classrooms suggest a more uncommon claim: testing is a useful mnemonic device because the act of retrieving information from memory has a direct effect on the later retrievability of that information (see Karpicke & Roediger, 2007). Intriguingly, these direct effects, referred to as *retrieval practice* effects, are often robust, even without feedback or restudy opportunities (Carpenter, Pashler, & Vul, 2006; Hinze & Wiley, 2011; Roediger & Karpicke, 2006b).

Recently, research has focused on the more difficult questions of *when* and *why* retrieval practice may enhance retention (Carpenter, 2009; Pyc & Rawson, 2010), and *when* and *why* retrieval practice may influence comprehension or conceptual understanding, as demonstrated on final tests requiring transfer (Butler, 2010; Johnson & Mayer, 2009; Karpicke & Blunt, 2011). The current study addressed these questions specifically with regard to

* Corresponding author. Present address: School of Education and Social Policy and Department of Psychology, 2120 Campus Drive, Northwestern University, Evanston, IL 60208, United States.
E-mail address: s-hinze@northwestern.edu (S.R. Hinze).

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.03.002

whether the elaborative retrieval hypothesis can be applied to learning from complex scientific texts.

Elaborative retrieval and the testing effect

Some accounts of the testing effect focus on retrieval practice as the primary mechanism for enhanced retention (Roediger & Karpicke, 2006a). These accounts suggest that testing strengthens a memory trace by practicing the retrieval skills necessary for later retrieval. Karpicke and Blunt (2011) demonstrated that repeated free recall practice led to superior final test performance when compared to elaborative concept mapping and argued that retrieval likely reduces the number of cues used to retrieve an item from memory, rather than elaborating the connections between items. In this way, the power of retrieval is in strengthening the accessibility of individual memory traces, possibly by making those memory traces more distinctive (see also Karpicke & Smith, 2012).

In contrast, some researchers have suggested that retrieval practice can serve to elaborate the contents of mental representations (Carpenter, 2009; McDaniel & Masson, 1985) by encouraging the learner to re-organize or supplement initially encoded information. Carpenter (2009) proposed that one reason that retrieval benefits long-term memory over restudy is that retrieval is more likely to activate related elaborative information. This *elaborative retrieval hypothesis* is consistent with several pieces of data. Most convincingly, Carpenter (2011) demonstrated that cued-recall practice tests enhanced retention not only for target information (e.g. "child" in the pair "mother: child") but also enhanced retention for words with strong semantic associations with the pair (e.g. "father"). Enhanced retention of this "semantic mediator" suggests that retrieval attempts in this context served to broaden, rather than focus, the activation of information in semantic networks. The elaborative retrieval hypothesis is also consistent with data showing that more difficult retrieval attempts are more effective for long-term retention (Carpenter, 2009; Pyc & Rawson, 2010; see Bjork (1994) for a more general desirable difficulties framework). For instance, short answer tests are more demanding than multiple-choice tests, but are more effective for long-term retention (Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007). Similarly, more open-ended recall tests are more demanding and more effective than cued-recall tests (Glover, 1989; Hinze & Wiley, 2011). According to the elaborative retrieval account, more demanding retrieval attempts require the learner to actively reconstruct the content, and this reconstructive process necessitates the access of additional information that is then associated with the existing memory trace (Carpenter, 2011). This elaboration during retrieval is, at least in part, why those memory traces are more accessible at delayed tests.

Constructive retrieval and learning from texts

The goal of the present research is to consider what types elaborative processes during retrieval will enhance comprehension of complex text materials, with comprehension evidenced by enhanced transfer performance. (For our purposes, we define a "transfer" test broadly as any final test that differs from initial tests, either by testing the same materials in a different format, or by testing related but not identical information.) A few experiments on learning from text have demonstrated benefits of retrieval attempts, with feedback and/or restudy, on these sorts of transfer tests (Butler, 2010; Karpicke & Blunt, 2011; McDaniel, Howard, & Einstein, 2009). However, while more successful performance on transfer tests is thought to reflect better understanding of the text contents (Kintsch, 1994; Mayer, 2001; Wiley, Griffin, & Thiede, 2005), it is not clear what role, if any, *elaborative* retrieval played in obtaining these benefits (Karpicke & Blunt, 2011).

In order to apply the ideas of elaborative retrieval to complex learning situations, it helps to consider research and theory on text comprehension and learning from text. According to Kintsch (1994, 1998), learning from texts involves not only memory for words as presented (the surface form), but also the abstracted representation of propositions (the textbase) and a representation of the meaning of the text and its relationship to prior knowledge (the situation model). The situation model depends not only on the text itself, but on connections that are made between distal parts of the text and/or inferences based on prior knowledge. Because of this, building a coherent, enduring, representation of a text is a *constructive* process and is dependent on the generation of inferences during (or after) reading. Thus, to the extent that "elaboration" enhances long-term retention of textual materials, it may be through facilitation of these inferences with the aim of constructing a coherent representation of the text in memory.

This process of constructing coherent representations differs somewhat from the elaborative processes described in paired-associate learning (Carpenter, 2009, 2011). Consider the requirements of learning from science texts. While the types of associations elaborated in paired-associates learning may be facilitated through spreading activation, coherent situation model representations of science text content typically require a series of causal inferences to integrate pieces of information into an accurate mental model of the phenomena (Graesser, Leon, & Otero, 2002; Kintsch, 1994; Wiley et al., 2005). For example, readers of a text on cell mitosis need to not only remember the names of the phases (prophase, anaphase, etc.), but how one phase necessarily proceeds the previous step (and vice versa; see Millis & Graesser, 1994). Unfortunately, students with low prior knowledge, poor reading ability, or limited working memory capacity often have difficulty making these sorts of inferences while reading expository texts (e.g. McNamara, Kintsch, Songer, & Kintsch, 1996; Voss & Siflies, 1996; Wiley & Myers, 2003). Yet, measures of long-term learning tend to rely most heavily on situation-level representations, rather than rote memory of words or sentences (e.g. Kintsch, Welsch, Schmalhofer, & Zimny, 1990). Because of this, much research has focused on the conditions that may encourage the development of a coherent situation-model level representation, and more active or constructive processing during reading (Kintsch, 1998). It may be the case that these same sorts of *constructive* processes can be encouraged during retrieval practice, and may be in evidence

based on the quality of responses participants provide on practice tests.

Relatively little attention has been paid to differences in processing *within* retrieval tasks, so it is difficult to determine which conditions lead to more, or less, constructive processing during a retrieval attempt. While some researchers have analyzed differences in test formats (Kang, McDermott, & Roediger, 2007; McDaniel et al., 2007) or test scope (Hinze & Wiley, 2011), there may be vast differences in processing even within a test format. For example, participants may engage a broad or narrow search of memory in response to a cued-recall prompt (Chan, McDermott, & Roediger, 2006). This potential variance exists because recall is typically not a wholly automatic process. It is subject to control by a learner, and relies heavily on top-down, metacognitive control (Benjamin, 2008) and reconstructive processes. As a classic example, the content of recall can vary greatly depending on the participant's perspective at the time of recall (Anderson & Pichert, 1978). Thus, processing during retrieval attempts can differ based on a participant's individual goals or expectations, or based on the particular task demands of the test itself. And, the quality of this processing during testing may predict learning from text moreso than engaging in retrieval practice alone. We sought to manipulate just this type of qualitative processing difference, informed by research into manipulations that have been found to encourage more or less constructive processing of science text materials.

One major approach that has been used to encourage constructive processing during reading has been the manipulation of task instructions. For example, readers who are instructed to read for study engage in more inference construction during reading and subsequently demonstrate superior recall of text than readers instructed to read for entertainment (van den Broek, Lorch, Linderholm, & Gustafson, 2001). Alternatively, providing participants with example questions can alter expectations for learning outcomes and, in turn, influence processing and monitoring during reading (Thiede, Wiley, & Griffin, 2011). Specifically of interest for the current study, Thiede and colleagues found that participants who answered example questions requiring inferences were better able to monitor processes related to comprehension (inference-generation), than participants who answered example questions requiring only memory for details presented in the text. Finally, readers encouraged to construct *explanations* during reading tend to construct more coherent representations from texts than participants who are not encouraged to explain, leading to superior retention and learning outcomes (Chi, deLeeuw, Chiu, & Lavancher, 1994; Cote, Goldman, & Saul, 1998; Griffin, Wiley, & Thiede, 2008; McNamara, 2004).

The above studies used instructions that affected the constructive processing of material *during* reading. The present study examines whether these same kinds of instructions could have similar effects when presented *after* reading, during an open-ended practice test. If so, we would predict that engaging in processing aimed at constructing a coherent representation of the scientific content (i.e. constructive retrieval) should influence the efficacy of practice tests for long-term retention and comprehension outcomes. Specifically, we propose a *constructive retrieval hypothesis* whereby engaging in constructive processes during open-ended practice tests will lead to larger effects of these tests (relative to restudy) than engaging in processing focused on rote retrieval.

Overview of experiments

The current experiments explore the efficacy of the constructive retrieval hypothesis in accounting for learning from tests as measured by assessments of text comprehension. Generally participants read texts (encoding phase), then either restudied the material or engaged in practice tests under different conditions (practice test phase), followed by a 1 week delay and new tests of the material (final test phase). In all experiments, final tests were established to assess retention of both *details* and *inferences* derived from the texts (Kintsch, 1994). In Experiment 1 we describe these materials in more detail and provide baseline data on the final set of test items that will be useful for interpreting the results of Experiments 2 and 3.

In Experiments 2 and 3, we manipulated conditions during the practice test phase to encourage more, or less, constructive processing during practice tests. The manipulations were not aimed at altering the encoding phase, but rather directed at the practice test phase, as this was an exploration of the role of constructive processes *during* retrieval. In Experiment 2, processing was indirectly altered by giving participants example questions to influence expectations for learning outcomes. This manipulation allowed us to test whether readers would alter processing during retrieval based on the perceived relevance of different types of information, and how this may influence learning and retention. The constructive retrieval hypothesis predicts that focusing on inferences in the service of constructing a coherent response would be more effective for learning from retrieval attempts than focusing on specific details. In Experiment 3, free recall and explanation instructions were compared, with the prediction that encouraging explanations would lead to more constructive processing during practice tests and lead to superior learning and retention. This focus on the nature of processing during practice tests, and the products of that process, should help to clarify the role that constructive processes play in learning from tests.

Experiment 1

To examine whether retrieval practice can improve retention and comprehension, criterion tests were designed to assess both memory of specific details from a series of expository science texts and inferences that were not directly provided in the texts. This first experiment provides baseline performance data for the materials that were used in Experiments 2 and 3. Examples of detail and inference questions for one of the texts can be found in the Appendix. These items were designed according to principles consistent with Wiley et al.'s (2005) formulation to reflect text-based retention and situation-level comprehension. Text-based *detail* items could all be answered

Table 1
Example items and relevant text from the viruses reading.

Type of item	Relevant text	Question
Inference	Biologists consider viruses to be non-living because viruses are not cells . . . Viruses can only multiply when they are inside a living cell	According to information in the passage, which of the following could not be infected by a virus? A. Virus B. House plant C. Fungus D. Dog
Detail	The cells that viruses infect in order to multiply are called host cells.	The cells a virus infects are called: A. Support cells B. Lock cells C. Host cells D. Protein cells

using the information available in one sentence of the text, and were intended so that a simple search of memory, or of the text, if available, would yield the answer to the question. These questions were intended to reflect lower-level information from the text and not to require inferences beyond the information that was given. *Inference* items were designed to require the integration of information from multiple sentences since the answer to the question was not directly stated in the text. These items may require the learner to predict what might happen in a novel situation, to select the appropriate explanation for a phenomenon, or to select the correct order of events in a scientific process. Thus, these *inference* items were intended to require the types of inferences that are necessary for situation-level comprehension of science texts (Graesser et al., 2002; Kintsch, 1994; Wiley et al., 2005).

We present these baseline data because it is important to establish the degree to which our measurements of memory for details and inferences are sensitive to participants' attempts to comprehend the text materials. In other words, we hope to determine the functional ceiling and floor levels of performance for these materials. A reasonable estimate of ceiling performance can be conceptualized in two ways: performance on the criterion test questions with the text available, and performance on the criterion test immediately after reading. The former provides an estimate of expected performance given perfect memory for the surface form of the text, while the latter provides an estimate of expected performance given typical levels of initial encoding. Since retrieval practice effects may work in part by attenuating the rate of forgetting (Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003), it is especially important to know what the relative levels of performance would be before forgetting is allowed to set in.

Method

Participants

Introductory psychology students were randomly assigned to one of three conditions: prior knowledge ($n = 36$), text available ($n = 33$), or immediate test ($n = 34$).

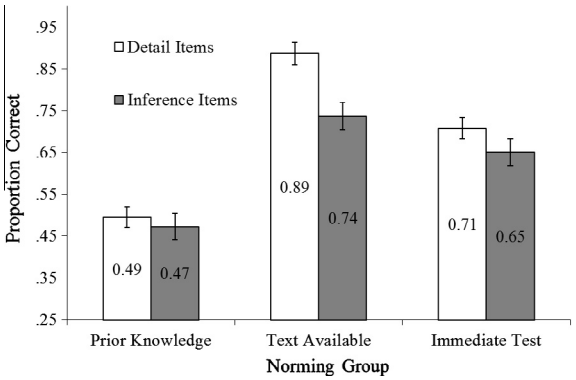

Fig. 1. Baseline data on critical texts for Prior Knowledge, Text Available, and Immediate Test conditions on Detail and Inference Test Items. Chance accuracy was .25.

Materials and procedure

Five short expository 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: the endocrine system, vision, the respiratory system, viruses, and metabolism. The texts ranged in length from 335 to 432 words with Flesch reading ease ratings from 50.5 to 70.7 and Flesch–Kincaid grade level between 7.0 and 9.9 and were presented as a whole without pictures or diagrams.

The tests consisted of multiple choice questions intended to assess retention and comprehension of the texts. Five *detail* questions and five *inference* questions were created for each text (see Table 1 for an example of each, and the Appendix for full materials from one reading). For all questions, there was only one correct answer.

The prior knowledge baseline group simply received the test questions without ever reading the text and were asked to answer based on their prior knowledge. The text available group was given the texts along with the tests and was told to refer to the texts to answer the questions.[1] The immediate test group read through each of the texts twice and then was asked to answer the test questions from memory.

Results and discussion

We present the data only for three critical texts (chosen for later use in Experiments 2 and 3) that demonstrated the greatest difference between prior knowledge and text available groups. We selected these texts based on reasoning that studying these materials would have the greatest value for learning beyond prior knowledge. These critical

[1] The text-available group was not intended to replicate an open-book testing scenario where participants may read the text before responding with the text available. For an exploration of learning from open and closed-book tests see Agarwal, Karpicke, Kang, Roediger, and McDermott (2008). Rather, our intention was to optimize the availability of the surface features of the text, which could be done with or without initial reading. Future studies may benefit from consideration of a variety of baseline conditions, including those utilized here and others such as open-book testing.

texts were on the topics of vision, viruses, and the respiratory system. All of the analyses presented below remain significant if all 5 texts are included in the analysis.

Fig. 1 shows mean performance for the critical texts for the prior knowledge, text available and immediate test conditions on both detail and inference items.

The baseline performance data indicate that the participants were able to answer nearly half of the items based on prior knowledge of these topics. However, immediate test performance levels were clearly higher than prior knowledge ($t(68) = 5.06$, $p < .001$), indicating that there was substantial initial learning and that participants *did* find reading the texts valuable. Immediate test performance was lower overall than with text available ($t(65) = 3.91$, $p < .001$), indicating that participants did not encode the full extent of the texts.

Inference performance was lower than detail performance with text available ($t(32) = 6.12$, $p < .001$) and after an immediate test ($t(33) = 2.35$, $p = .03$). This may not be surprising given that the inference questions were designed to assess deeper levels of comprehension that may not be as easily ascertained in a search task (text available) or after simply re-reading a text (immediate test). Note that this difference between detail and inference content suggests that these outcomes could be considered to be on different scales. One implication of this for the following experiments is that equivalent absolute improvement (relative to control) on detail and inference tests may actually reflect a larger relative gain for inference content than for detail content.

These data show that reading the texts had value for both detail and inference items, and that there was room for improvement relative to prior knowledge alone. With these baselines established, we next turned to manipulations designed to encourage more or less constructive retrieval during testing, and explore the influence of these processes on long-term retention and understanding.

Experiment 2

In this experiment, we attempted to alter learners' processing during retrieval attempts by providing example test questions prior to retrieval practice. Expectancies derived from these example questions for either detail or inference final tests were intended to alter retrieval processes, relative to retrieving with no expectancy. To the extent that engaging in retrieval practice alone should increase learning, recall with no expectancy should lead to superior performance over rereading (a retrieval practice effect). The more interesting question was whether and how the example questions would influence processing during retrieval. We predicted that example test questions, presented after reading, would influence the type of information that is generated during a recall attempt. For instance, after receiving example detail questions, participants may generate the kinds of detailed surface information appropriate for detailed tests. In contrast, after receiving example inference items, participants may attempt to construct a more coherent response aimed at the kinds of connections between ideas relevant for inference tests. Thus, our first prediction was that providing participants with example questions would *focus* retrieval attempts, reducing the total amount of retrieval on practice tests relative to recalling without example questions. If this result holds, it is possible that both expectancy conditions might show reduced testing effects relative to recall with no expectancy, simply because less information is generated at practice. However, the *constructive retrieval hypothesis* predicts that an inference expectancy may lead to more constructive processing during the recall attempt, which should more effectively facilitate learning relative to a detail expectancy.

Method

Participants

Ninety-seven introductory psychology students completed all parts of the experiment for course credit, after eliminating seven participants with incomplete data (i.e. they did not respond to all initial or final test questions). Participants were randomly assigned in a between-participant design to each of the four groups defined in Table 2 below.

Materials

The texts normed in Experiment 1 were used, lettered A–E in Table 2. The texts from which example questions were derived are texts A and B (metabolism, the endocrine system). The critical texts for which the Initial Study/Test manipulation varied were C, D and E (vision, viruses, and respiratory system). The final detail and inference test questions were also identical to those normed in Experiment 1.

Procedure

All aspects of the task were presented on computers in small groups and were self-paced with no time limit. Recall attempts were written on paper to allow for non-sequential responses. Participants read all five texts in succession in the same order (metabolism, endocrine system, vision, viruses and respiratory system). Each text was presented along with a title on a single screen with no need to scroll

Table 2
Outline of conditions for Experiment 2.

Condition	Encoding	Test expectancy	Initial Study/Test	Final MC tests (counterbalanced)
Reread (control)	Read texts A–E	Re-Study A and B	Reread C–E	Detail A–E; Inference A–E
Paragraph recall	Read texts A–E	Re-Study A and B	Recall C–E	Detail A–E; Inference A–E
Paragraph recall-expect detail	Read texts A–E	Example Detail questions A & B	Recall C–E	Detail A–E; Inference A–E
Paragraph recall-expect inference	Read texts A–E	Example Inference questions A and B	Recall C–E	Detail A–E; Inference A–E

(Sanchez & Wiley, 2011). Following initial reading, procedures differed based on condition.

Example questions. The first two non-critical texts (on metabolism and the endocrine system) were used to present example questions for some participants after initial reading. The Paragraph Recall (n = 25) and Reread (n = 26) groups received no example questions and instead simply reread the non-critical texts in order. The other two groups answered either five detail questions for each non-critical text (n = 23) or five inference questions for each non-critical text (n = 23). Both groups were informed that they would practice multiple-choice tests. Those in the Expect Detail group were informed that, "These tests are designed to test your memory for specific details from the texts" while those in the Expect Inference group were informed that, "These questions test your comprehension, or your ability to make connections across parts of the text." Both groups then received the example detail or inference questions for both non-critical texts.

A manipulation check was then conducted with a test expectancy rating scale to establish that the example questions task was effective at instilling test expectations. The two example question groups rated their agreement with a series of statements on a scale from 1 to 4. Two questions were coded as detail expectancy (e.g. "The answer to these questions could be found in a single sentence from the reading.") while another two questions were coded as inference expectancy and were reverse scored (e.g. "I had to use reasoning to find the answers to these questions, not just my memory.").

Initial study/test exercises. After completing the example questions (or rereading) for the non-critical texts, participants either reread or recalled the next three critical texts. The recall instructions asked participants to use topic sentence cues to recall the relevant content from each paragraph, but that they could use their own words. Participants who received example questions for the non-critical texts were also asked to "recall information that will be useful for tests of the type you practiced." For each recall attempt, participants wrote their responses on paper with topic sentence cues (see Appendix A), and pressed a button on the keyboard when finished in order to record time on task.

Delay. After completion of all tasks in Session I, all participants were dismissed and returned for Session II after a 7 day delay.

Final tests. Upon returning for Session II, participants completed multiple-choice tests for all five readings in the same order as the original presentation. While the order of the topics was held constant, the order of detail and inference questions was counterbalanced.

Design

For the three critical readings (vision, viruses, and the respiratory system) the design was a 4 (Initial Study/Test; reread, paragraph recall, paragraph recall-expect detail, paragraph recall-expect inference) by 2 (Final Test; detail, inference) mixed design, with Final Test as a within-participant variable and Initial Study/Test varying between-participant.

Results

Manipulation check

To assess whether participants in the example question groups had different test expectancies, we examined average scores of the expectancy rating scale described earlier. Participants who answered example detail questions rated the example questions more consistent with detail expectancy (M = 2.92, SD = .52) than participants who answered example inference questions (M = 2.29, SD = .51) [$t(41) = 3.99$, $p < .001$, $d = 1.21$].

Final test performance

The effects of Initial Study/Test activities were compared on the two types of Final Tests collapsed across the three critical readings. We observed no interpretable relationship based on the order of the texts (cf. Wissman, Rawson, & Pyc, 2011). Performance was highest for the second text, but this may simply be because the content (viral

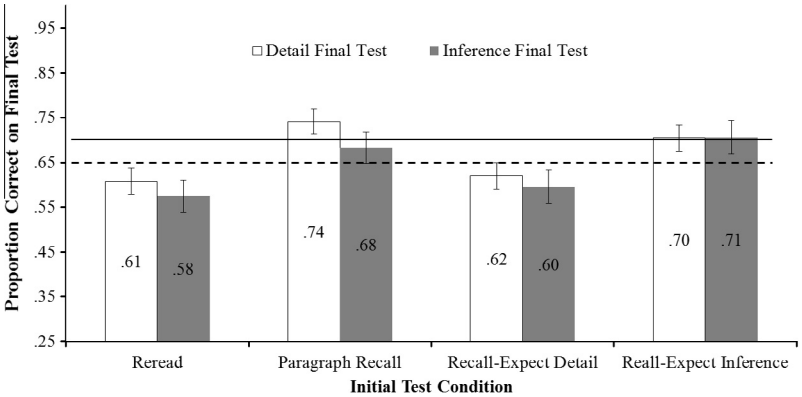

Fig. 2. Effect of initial study/test activity on final test performance. Error bars represent ±1 SEM. Chance accuracy was .25. Norming data from Experiment 1 for immediate test performance is represented by the solid (detail items) and dashed (inference items) horizontal lines.

replication) was easier for students to understand, rather than any effects of recall order. The data relevant for the overall analyses are shown in Fig. 2.

Performance on detail items was slightly higher overall (M = 67%, SD = 15%) than for inference items (M = 62%, SD = 18%) as indicated by a main effect of Final Test, $F(1, 93) = 4.17$, $p = .04$, $\eta_p^2 = .04$. Importantly, there was also a main effect of Initial Study/Test, $F(3,93) = 4.58$, $p = .01$, $\eta_p^2 = .13$. Following up the main effect of Initial Study/Test, the paragraph recall group outperformed rereading, $t(49) = 2.68$, $p = .01$, $d = .75$, demonstrating a testing effect. Rereading was also outperformed by the expect inference group, $t(46) = 2.67$, $p = .01$, $d = .78$. In contrast to the other tested groups, the expect detail group performed relatively poorly, with significantly lower overall performance compared to the paragraph recall, $t(47) = 2.50$, $p = .02$, $d = .72$, and expect inference groups, $t(44) = 2.56$, $p = .01$, $d = .75$. No other comparisons were significant (ts < .36).

There was no significant interaction between Initial Study/Test condition and type of Final Test. That is, the benefits of paragraph recall and expect inference activities were equally robust for detail and inference test items, $F(3,93) = .76$, *ns*.

In sum, both the paragraph recall group and expect inference group outperformed the rereading group, demonstrating a testing effect on both detail and inference items.[2] In addition, the example questions played a role in influencing the effectiveness of recall practice tests, as the group with example detail questions performed as poorly as rereading on both types of final test questions.

Protocol analyses

Given that three groups of participants had similar recall tasks, but differed substantially in performance on final tests, it stands to reason that the quality and/or quantity of processing during recall attempts were responsible for these differences.

Word counts and time on task. In order to establish a measure of the quantity of recall in a given protocol we calculated a simple word count for each of the three responses and obtained a mean for each participant. We predicted that the example questions would focus retrieval, thus reducing the total amount of retrieval and/or time on task. As expected, the paragraph recall group with no example questions provided substantially more content (M = 86.64, SD = 35.52) than expect inference (M = 62.09, SD = 26.44; $t(47) = 2.71$, $p = .01$) or expect detail (M = 53.68, SD = 62.09; $t(47) = 3.88$, $p < .001$). The two example questions groups did not differ, $t(44) = 1.19$, *ns*. Similar patterns were found for time-on-task with the paragraph recall group spending more time recalling (M = 326 s, SD = 86 s) than the groups with example questions (expect detail, M = 225 s, SD = 75 s; expect inference, M = 223 s, SD = 59 s). It appears that the example questions did focus retrieval, limiting the amount of information produced and the time spent retrieving.

Explanation quality coding. The predicted processing differences based on test expectations were qualitative, rather than quantitative in nature. To code for the quality of the content included in the protocols, we created a holistic explanation quality scale from 1 (Very Poor) to 5 (Very Good) intended to be similar to the scoring that would be obtained if an educator were to analyze these responses for summative or formative evaluation purposes (see Lehman, Schraw, McCrudden, & Hartley, 2007, for a similar coding scheme). Two coders, including the first author, coded all responses, blind to condition, on the degree to which the response provided a coherent explanation for the scientific process described in the text (e.g. how light passes through the eye and how light is regulated, manipulated, and interpreted on this path). This rating was made for all three responses for each participant, so that an average explanation quality score could be obtained for each participant. Based on Gamma correlations the two coders were highly reliable on evaluating the responses for the critical vision, viruses, and respiratory system texts (γ's = .95, .91, .90 respectively). Responses that received the lowest explanation quality scores typically contained bullet-pointed lists of information that were not tied together in any coherent manner. Responses receiving moderate scores contained internally coherent responses within paragraphs, but did not explicitly tie these parts of the process together into a coherent whole. Responses that received the highest explanation quality scores focused on the order of steps communicated by the text (all texts described a sequential process) and described how one step necessarily led to each of the following steps.

If the quality of the explanations were related to long-term retention following the initial test phase, then explanation quality scores should be related to final test performance. This was the case, as average quality scores correlated significantly with final detail and inference test scores, as well as overall performance (r's = .46, .67, and .63 respectively, all p's < .001). More importantly, we were interested in whether explanation quality differed based on test expectations. In terms of overall means, we observed a main effect of Initial Study/Test condition ($F(2,72) = 5.56$, $p = .01$, $\eta_p^2 = .14$), whereby explanation quality was higher in the paragraph recall (M = 2.92, SD = .99) and expect-inference (M = 3.06, SD = 1.04) groups than expect-detail (M = 2.19, SD = .82). However, qualifying these results for overall means, we found that recall quality was positively correlated with response length, as measured by word counts (r = .68, p < .001). In order to compare the quality of responses *independent* of response length, we conducted an ANCOVA comparing response

[2] Another option for quantifying the effects of the rereading and testing conditions is to scale final test scores based on the estimated baseline and ceiling performance levels observed in Experiment 1. In this analysis, the range of possible scores for detail tests would be .40 (.89–.49) and the range for inference scores would be .27 (.74–.47). Converting to these scales gives an estimate of the amount of learning retained given a more reasonable range of expected scores. In this scheme, retention of detail learning would be .30 for reread, .63 for paragraph recall, .33 for expect detail, and .53 for expect inference. The values for retention of inference learning would be .41 for reread, .78 for paragraph recall, .48 for expect detail, and .89 for expect inference. This analysis demonstrates that final test performance was quite strong in the paragraph recall and expect inference groups relative to rereading and expect memory groups with relatively strong learning gains and retention. We thank Henry L. Roediger, III for suggesting this analysis.

quality for the three recall groups, controlling for average word counts. In this analysis, mean explanation quality was similar for paragraph recall ($M = 2.49$) and expect-detail ($M = 2.53$) groups, and was highest for the expect-inference group ($M = 3.20$). The main effect of condition was significant after controlling for word counts, $F(2,68) = 7.79$, $p = .001$, $\eta_p^2 = .19$. These analyses demonstrate that explanation quality was related to final test performance and was highest in the groups with high final test performance: paragraph recall and expect-inference. But the high scores for paragraph recall were due, in large part, to the sheer quantity of information recalled with no test expectation.

Discussion

The results of Experiment 2 suggest that practice recall tasks can have an influence on learning from text as measured with both detail and inference items. These results suggest an influence of retrieval practice not just on memory for surface features, but also on understanding of scientific concepts (Kintsch, 1994; Pellegrino, 2012; Wiley et al., 2005). Interestingly, performance levels for the successful groups were as high as (or nominally higher than) performance in the Immediate Test comparison condition from Experiment 1, as indicated by the horizontal lines in Fig. 2 ($M = .71$ for detail items, $M = .65$ for inference items), suggesting little forgetting over the 7-day delay in the two successful tested groups. Clearly, these data demonstrate the robustness of retrieval attempts for long-term learning.

This experiment also demonstrated that the types of information generated during retrieval attempts may underlie the benefits of these interventions. Participants who received example test questions provided less content in their practice recalls, presumably because they were attempting to target their retrievals toward the appropriate content. Importantly, when participants expected a detailed test, the responses were relatively incoherent, and retrieval practice was no more effective than rereading. In contrast, participants were able to put the inference example questions to good use in constructing coherent responses aimed at inference-relevant content, and subsequently performed well on the final transfer tests.

Thus, it appears that retrieval attempts may be effective tools for learning when the conditions of retrieval encourage constructive processing aimed at coherent explanatory responses. However, several issues preclude strong support for this conclusion. The expect inference group did not outperform the paragraph recall group overall. One possible reason for this is that retrieval without example questions naturally involves this sort of constructive processing. In fact, the paragraph recall group demonstrated high scores on explanation quality. However, we also found that these high scores were dependent on the sheer length of responses in the paragraph recall group. This suggests the possibility that retrieval practice effects can be enhanced either by increasing the quantity of content retrieved (no example questions) or the constructive quality of that information (inference example questions). Further support for the role of explanation quality, independent of response length, may help clarify this point. Another issue is that the topic sentence cues in the paragraph recall task may have provided an organizational structure that supported coherent responses (except in the expect-memory group). The role of constructive processing during recall may be even stronger or more obvious when participants are required to construct their own organizational scheme. Experiment 3 was designed to address these issues.

Experiment 3

Experiment 2 showed that participants infer different expectations for learning outcomes from example questions and can bias their recall content based on these expectations. The findings suggest that constructive, explanatory responses may have been particularly effective or efficient for long-term learning. This experiment more directly manipulated constructive processing during retrieval by comparing free recall against a condition where explanation was explicitly prompted as part of the recall task. The explanation prompt was added in light of research on the benefits of engaging in explanation activities for understanding expository texts, including difficult science texts (Chi et al., 1994; Cote et al., 1998; Griffin et al., 2008; McNamara, 2004). These studies suggest that explanation attempts prompt readers to construct inferences that move beyond text content by explicitly instructing readers to make connections between ideas. This switch between implicit demands of instilling test expectancies in Experiment 2, to the explicit demands of the explanation instruction in Experiment 3, represents the main difference between the two experiments.

A smaller change that was required to better assess the importance of constructive processing was that the topic sentence cues used for recall of the critical texts in Experiment 2 were removed. While these cues may serve as a useful scaffold for retrieval, they may also limit the importance of global coherence-building processes, such as reorganization, thematic generalization, or integration across distant ideas that might occur in more open-ended recall tasks (Kintsch, 2005; Zaromb & Roediger, 2010). Not only may the cued recall format prevent such constructive processing from occurring by prompting the retrieval of particular paragraphs, it certainly limits the ability to *assess* the use of constructive processes during recall, since responding to the prompts constrains the scope, content, and organization of the recall. Thus, in this Experiment, open-ended free recall was employed to allow for a wider range of processes and strategies to be observed in the recall protocols.

While explanation instructions are typically given before reading as a type of encoding manipulation, the present study examined whether giving the same kind of instruction may also improve learning when presented during an open-ended practice test, and *after* reading. To the extent that engaging in retrieval practice enhances learning, unguided free recall should lead to superior performance over rereading (a *retrieval practice* effect). However, the *constructive retrieval hypothesis* predicts that, to the extent that explanation instructions lead to more constructive processing during recall, retrieval in the

context of explanation instructions should lead to superior learning relative to unguided free recall.

Method

Participants

Ninety-three introductory psychology students completed all parts of the experiment for course credit and were included in the analyses below. Additionally, five participants were eliminated due to incomplete data during initial tasks and two participants were excluded for non-compliant behavior (one participant was evidently intoxicated; another admitted to guessing randomly on all final test responses).

Materials and procedure

The texts were identical to those used in Experiments 1 and 2.

Initial study/test exercises. All participants read each of the five texts in the same order on a computer at their own pace. Immediately afterward, all participants reread the first two topics (metabolism and the endocrine system). The reason for retaining these two texts in the procedure was so that the repeated exposures to the critical texts would be similar in spacing and timing to that of Experiment 2. Then, three groups varied in a between-participants design on Initial Study/Test exercises ($n = 31$ for each group). The instructions that appeared for each of the three critical texts are provided below:

- *Reread*: Practice reading the [vision] text again. Take your time reading as your time on task WILL be recorded.
- *Free recall*: Practice retrieving the content from the [vision] text. You may use your own words or those from the text. Your response will be scored on how much of the text you can recall.
- *Explain*: Practice writing an explanation of the [vision] text. Use your own words to communicate the text's explanation for how [vision] works. Your response will be scored on how completely and accurately you can explain the topic.

Recall attempts were written on paper to allow for non-linear response strategies and organization. All groups pressed a button on the computer when they completed each task to get an estimate of their time on task. All readings and activities were presented in the same order for all participants.

Delay. After completion of all tasks in Session I, participants were dismissed and returned for Session II after a 7 day delay.

Final detail and inference tests. These were identical to the test items normed in Experiment 1 and used in Experiment 2. Order of presentation was counterbalanced as in Experiment 2.

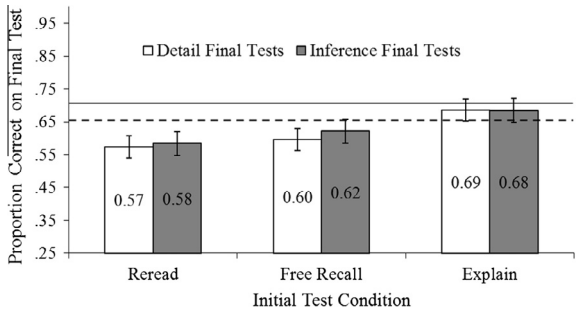

Fig. 3. Effect of initial study/test activity on final test performance in Experiment 3. Error bars represent ±1 SEM. Chance accuracy was .25. Norming data from Experiment 1 for immediate test performance is represented by the solid (detail items) and dashed (inference items) horizontal lines.

Design

This experiment was a 3 (Initial Study/Test: reread, free recall, explain) by 2 (Final Test Type: detail, inference) mixed design with Final Test Type varying within participants and Initial Study/Test varying between participants.

Results

Final test performance

As shown in Fig. 3, performance on detail items was roughly equal overall ($M = 61.9\%$, $SD = 19.2\%$) to inference items ($M = 63.0\%$, $SD = 20.5\%$) and there was no main effect of Final Test Type, $F(1,90) < 1$. Similar to Experiment 2, there was no interaction between Final Test Type and Initial Study/Test indicating that any effects of Initial Study/Test were equivalent for detail and inference items, $F(2,90) < 1$. Note, however, that the inference items have a lower ceiling than detail items based on the norming from Experiment 1, so learning of inference content may have been particularly robust.[3]

There was a main effect of the Initial Study/Test manipulation, $F(2,90) = 3.02$, $p = .05$, $\eta_p^2 = .06$. Following up the main effect, the free recall task did not improve performance over reread, $t(60) = .64$, *ns*. The difference between explain and reread was significant based on a one-tailed test, $t(60) = 2.31$, $p = .01$, $d = .59$. Finally, the difference between free recall and explain groups was significant based on a one-tailed test, $t(60) = 1.87$, $p = .035$, $d = .47$. Performance in the explain group was as high (or nominally higher in the case of Inference items) as the Immediate Test norming group from Experiment 1, as indicated by the horizontal lines in Fig. 3 ($M = .71$ for detail items, $M = .65$ for inference items).

Protocol analyses

Given that the explanation instructions were intended to alter processing during retrieval practice, it is important

[3] We also used the scaling procedure from the note from Experiment 2 to scale retention relative to baseline and ceiling estimates. Retention of detail learning can be estimated at .20 for reread, .28 for free recall, and .50 for explain. Retention of inference learning can be estimated at .41 for reread, .56 for free recall, and .81 for explain. This analysis demonstrates that learning was rather robust, especially in the explain condition.

to establish whether there were differences in the quantity and/or quality of initial test responses within and between conditions. In addition to quantitative measures of protocol length and the proportions of test-relevant content that were included, more global assessments were pursued to assess qualitative differences in the protocols.

Word counts and time on task. Time on task was roughly equivalent across tested groups with an average of 275 s per text for the free recall group (SD = 110) and 283 s per text for the explain group (SD = 93), $t(57) = .32$.

Average word counts for all three responses were used as a measure of the quantity of recall. The free recall group (M = 76.26, SD = 48.66) and explain group (M = 84.44, SD = 32.43) provided similar amounts of recall, $t(60) = .78$, *ns*.

Explanation quality coding

Despite similar amount of response content after free recall and explanation instructions, there may be qualitative differences in these responses that account for the more robust effects of explanation practice relative to free recall practice. Providing an explanation should require attempts to construct a coherent representation of the phenomenon described in each text, including causal inferences.

As with Experiment 2, two coders, including the first author, coded all responses, blind to condition, on the degree to which the response provided a coherent explanation for the scientific process described in the text. Based on Gamma correlations the two coders were highly reliable on evaluating the responses for the critical vision, viruses, and respiratory system texts (γ's = .89, .80, .88 respectively). If the explanation instruction was effective, then the quality of responses should be higher in the explain group than in the free recall group. This was the case, with higher average explanation quality scores after explanation instructions (M = 3.42, SD = .99) than after free recall instructions (M = 2.65, SD = 1.16) [$t(60) = 7.99$, $p = .01$, $d = .71$]. Also, if constructive processes are useful for learning during the initial test phase, then explanation quality scores, which reflect these constructive processes, should be related to final test performance. This was the case as well, as average quality scores correlated strongly with final detail and inference test scores, as well as overall performance, r's = .61, .58, and .67 respectively, all p's < .001.

If explanation instructions led to higher quality explanations than free recall instructions, and higher quality explanations were related to greater performance on final tests, then the quality of the responses may mediate the benefits of explanation-based testing on learning. Following this line of reasoning, participants in the free recall condition may perform well if they spontaneously construct coherent explanations during recall, and participants in the explain condition may perform poorly if their responses are more like those seen in free recall. Indeed, Explanation Quality was related to overall final test performance in both the free recall group ($r = .59$, $p < .001$) and the explain group ($r = .71$, $p < .001$).

To test for the mediating role of constructive processes during initial tests on learning more directly, we conducted multiple regression analyses with overall final test performance as the outcome variable, and Initial Study/Test Condition (free recall versus explain), Explanation Quality Score, and Word Counts (to control for recall length) as predictors. Explanation Quality scores were a significant predictor of final test performance when entered alone [$\beta = .67$, $t = 6.95$, $p < 001$]. This effect remained significant even when controlling for Word Counts and Explanation Condition [$\beta = .66$, $t = 3.79$, $p < 001$]. Interestingly, while the explain group showed higher performance than the free recall group when Initial Study/Test Condition was entered along with Word Counts [$\beta = .18$, $t = 1.69$, $p = .048$, one-tailed], this relationship was eliminated when Explanation Quality was included in the model [$\beta = .07$, $t < .07$]. Thus, it appears that the quality of the initial retrieval attempts related strongly to final test performance, a relationship that mediated the stronger testing effect when participants were encouraged to explain during retrieval, compared to when they were asked simply to recall.

Discussion

The main finding of this experiment, based on both final test performance and the protocol analyses, was that a testing effect was found when readers engaged in explanation during open-ended practice tests. This result is consistent with the constructive retrieval hypothesis. While an explanation instruction was effective at facilitating final test performance relative to rereading, no benefit was seen for free recall over rereading, suggesting that retrieval practice attempts alone cannot explain the benefits from explanation testing seen here.

Analyses of the written protocols further support the notion that constructive processes were responsible for the testing effects observed in this study. Regardless of condition, protocols demonstrating more coherent representations of the phenomena were associated with higher final test performance. Importantly, while some participants in the free recall group provided good explanations, and some participants in the explain group did not, this simple instructional manipulation was effective at fostering higher quality explanations. This effect may be particularly impressive given that participants were not given any training in our study on how to write an explanation, an issue that often limits the effectiveness of self-explanation manipulations on improving learning outcomes (see McNamara, 2004). An interesting question for future studies is whether providing training in explanation would lead to even larger advantages of explanation-based retrieval attempts.

General discussion

These experiments generally support the notion that engaging in retrieval of texts can be a robust mnemonic technique, demonstrating several instances where initial retrieval attempts aided future performance relative to rereading controls (and no instances where rereading outperformed retrieval attempts). These positive effects serve

to replicate the retrieval practice effects that are prevalent in the literature (Roediger & Karpicke, 2006a). In addition, where benefits for testing were found, they serve as evidence of transfer as final tests always assessed learning using novel questions not previously encountered. These transfer effects contribute to emerging data demonstrating the robustness of retrieval practice as a meaningful learning event. In fact, while there is some evidence that retrieval practice can improve performance on transfer tests when learning from text (Butler, 2010; Karpicke & Blunt, 2011; McDaniel et al., 2007, 2009), the present transfer effects are among the first that have been obtained in the absence of providing feedback or restudy opportunities. While feedback is clearly useful following practice tests (e.g. Butler & Roediger, 2008), the current effects demonstrate the robustness of the effects of the tests alone.

Yet, one theoretical and empirical extension shown here was the demonstration that not all retrieval attempts are equal, even within open-ended retrieval tasks, and that differences in processing during retrieval can greatly influence the effectiveness of such activities. In both experiments, we demonstrated that the *quality* of retrieval during initial tests could be influenced by the types of expectations or instructions that participants were given. These processing differences during initial tests, in turn, had downstream effects on learning and retention. These findings suggest that the influence of testing events on learning from text is complicated not just by differences between question types like multiple-choice vs. short answer questions (Duchastel & Nungester, 1982; Kang, et al., 2007), or cued vs. free recall prompts (Glover, 1989; Hinze & Wiley, 2011), but also by the different levels of representation inherently involved in learning from texts (Kintsch, 1998). When comprehending science texts, not only do readers need to encode what the text said, but they also need to construct an integrated representation of what the text meant including causal and other relationships between ideas (Graesser et al., 2002; Kintsch, 1994; Wiley et al., 2005). Clearly not all testing activities will encourage this latter type of constructive processing. Specifically, we found that neither retrieval in preparation for a test requiring memory for specific details (Experiment 2) nor retrieval in the context of unsupported free recall instructions (Experiment 3) led to learning gains from the testing events. Only conditions in which initial retrieval attempts focused on the information relevant for constructing coherent explanations of the phenomena led to robust benefits of testing over rereading.

One interesting finding is that we did not observe a differentiation in testing effects for detail versus inference test outcomes. Based on transfer-appropriate processing theory (Morris, Bransford, & Franks, 1977), it may be somewhat surprising that presenting example detail items in Experiment 2 failed to provide specific benefits for final tests with detail items. While this sort of pattern was certainly possible (see Johnson & Mayer, 2009), there are a number of reasons why such an expectation may not have been beneficial, even for detail tests. First, not all details from the texts were assessed on the final tests. So there was no guarantee that all details that were retrieved during the initial test would match those that were assessed on the final test. Second, as indicated by the norming in Experiment 1, *immediate* test performance was relatively high for detail items. This suggests that example detail items may be perceived as easy or trivial, thus encouraging less effortful or constructive retrieval, which would not be beneficial for long-term learning (see Pyc & Rawson, 2010). Finally, the general benefits for constructive processing on *both* detail and inference items is fully consistent with theories of text comprehension (Kintsch, 1994). Constructing a coherent situation model based on the text content should facilitate retrieval not only of inferences, but also facilitate an efficient search through memory for more specific details.

We also observed an interesting pattern of results across Experiments 2 and 3, where recall cued by topic sentences (E2; paragraph recall) resulted in a retrieval practice advantage over rereading, whereas free recall (E3) did not. We are hesitant to make strong conclusions given the cross-experiment nature of these differences. However, there are a few reasons why paragraph recall may have been more effective than free recall. First, the cues may have scaffolded initial recall performance. This interpretation is supported by the findings of longer recall responses for paragraph recall (E2: $M = 86.64$, $SD = 35.52$) than free recall (E3; $M = 76.26$, $SD = 48.66$). Additionally, the topic sentences may have served as a useful organizational tool, allowing participants to structure recall around the relatively sequential series of paragraphs. This interpretation is less well supported given that explanation quality scores were only a small amount higher for paragraph recall ($M = 2.92$, $SD = .99$) than free recall ($M = 2.65$, $SD = 1.16$), despite the longer recalls. The question of how much scaffolding is necessary to facilitate effective retrieval practice, without becoming overly specified (Hinze & Wiley, 2011), is an interesting one for future experiments.

Based on the data presented here, it appears that the benefits of initial retrieval attempts on learning from text cannot be limited to the benefits of retrieval practice alone, but rather seem to rely on constructive processes that may occur during retrieval attempts. Some types of testing may alter the organization of information available in episodic memory, allowing for fragmented representations to become more integrated. Such an explanation of how testing may alter learning from text is consistent with other results from studies on memory for categorized lists. Zaromb and Roediger (2010) found that increases in measures of list organization across sequential retrieval attempts accompanied the increased retrievability of the individual items. Thus, constructive processes aimed at constructing coherent explanations may serve to alter the structure or organization of the retrieved information, which allows for more efficient retrieval during final tests.

An additional way of viewing the benefits of constructive processes during initial retrieval attempts can be made in text processing terms. Engaging in constructive processes during a retrieval attempt may prompt the recognition of connections or the generation of new inferences that were not constructed during initial reading. For instance, a reader may not have originally inferred that the amount of light that reaches the retina is regulated by the iris. However, when asked to explain how light passes

through the eye, the participant may construct this inference based on their retrieval of two previously unconnected pieces of information: that the iris surrounds the pupil and that the pupil changes size to allow in more or less light. Rather than just a re-organization of the information, under such a scenario learners use their prior knowledge about causal relationships to construct and add new inferences into an existing, incomplete, representation. This constructive activity would lead to a more coherent model of the subject matter, which would allow for more efficacious retrieval of information during final tests.

The constructive retrieval hypothesis serves as an instantiation of the elaborative retrieval hypothesis (Carpenter, 2009, 2011) for complex text materials. While the elaborative retrieval hypothesis has been explored with regard to semantic associations in paired-associates learning, the types of "elaborative" processing necessary for learning from text (i.e. coherence-building, explanation) may be informed by the text comprehension literature (e.g. Cote et al., 1998; Graesser et al., 2002; Kintsch, 1994, 1998; Millis & Graesser, 1994). Karpicke and Blunt (2011), Karpicke and Smith (2012) and Karpicke and Zaromb (2010) have argued that these sorts "elaborative" mechanisms are not necessary to account for retrieval practice effects, suggesting the processes inherent to retrieval (e.g. discrimination of memory traces or strengthening of retrieval routes) are sufficient to account for retention and inference gains. It may certainly be the case that these discrimination and strengthening processes play a role in the fundamental advantage of retrieval practice over re-exposure. Our data, and those from other studies (Carpenter, 2009, 2011; Pyc & Rawson, 2010; Zaromb & Roediger, 2010) suggest that retrieval attempts *also* offer opportunities to reorganize and relate pieces of information in the course of constructing a goal-directed response. The constructive processes involved in some practice tests, or engaged by some participants, may have particular importance for the deep learning outcomes for the sorts of complex materials of interest in the current study, or in applied settings (see Pellegrino, 2012).

The constructive retrieval hypothesis predicts that processes that encourage the construction of coherent representations during encoding should have similar effects when engaged during a retrieval attempt. While the reconstructive nature of recall tasks means that learners *may* engage in constructive processing during any kind of open-ended retrieval attempt, evidence from these studies suggests that this kind of processing is not guaranteed and that testing events are more beneficial for learning from text when participants are directed toward doing it. The current studies specifically highlight that instilling inference-based test expectancies (Experiment 2) and prompting readers to construct explanations during practice tests (Experiment 3) can make testing events more effective as learning activities, specifically by encouraging the construction of coherent responses during practice tests. In conclusion, these experiments support a role for constructive retrieval in testing effects, and demonstrate that it is not retrieval practice alone, but rather the kind of constructive processing invoked during retrieval attempts that can improve both retention and comprehension when learning from text.

Acknowledgments

This research was supported in part by Grant R305B07460 from the Institute for Education Sciences Cognition and Student Learning to Jennifer Wiley. We wish to thank Drs. Katherine Rawson, Susan Goldman and Benjamin Storm for their help in the development of this project. We also wish to thank Thomas Griffin, Josh Redford, and Keith Thiede for their contributions, and Nicole Rivera, Aarti Sarup, Tim Hoskinson, Andrew Cho and Michelle Madison for their help with data collection and coding.

Appendix: Example materials from the vision text

Vision—Text

Please read the following text carefully. You may be tested on your comprehension of this text.

Your eyes are the sense organs that enable you to see the objects in your environment. Your eyes respond to the stimulus of light. They convert that stimulus into impulses that your brain interprets, enabling you to see.

Your eyes are made of many layers. The eyelid serves as the first line of defense, as closing it can protect your eyes from light and blinking can remove debris. The first visible layer of the eye includes the pupil and iris, but these are covered by the cornea—the clear tissue that covers the front of the eye. Light first passes through the cornea and enters the pupil. The pupil is the opening through which light enters the rest of the layers of the eye.

The iris is a circular structure that surrounds the pupil and can be many colors including brown, blue, grey and green. The iris regulates the amount of light entering the eye through the pupil. The size of the pupil is adjusted by muscles in the iris. The eye works best with a moderate amount of light is not too bright or too dim. In bright light, the pupil becomes smaller. In dim light, the pupil becomes larger.

After passing through the pupil, light reaches the lens. The lens is a flexible structure that focuses light. Muscles that attach to the lens adjust its shape. This bends the light rays producing an image that is in focus. To see things close up, the lenses in our eyes need to be thick to bend the light more. To see things far away, the lenses need to be thin to bend the light less. When people need glasses it is because their lenses aren't able to bend the light to the correct part of the eye by themselves.

After passing through the lens, the focused light rays pass through a clear, jellylike fluid in the middle of the eye. Then the light rays strike the retina, the layer of receptor cells that lines the back of the eye. The retina contains about 130 million receptor cells that each respond to light. There are two types of receptors. The cones work best in bright light and enable you to see colors and details when you look directly at something. They are mostly at the center of your retina. In contrast, rod cells work best in dim light and enable you to see black, white, and shades of

gray. Most of our retina is covered by rods, especially areas away from the center which is called the periphery.

When the rods and cones respond to light hitting them, nerve impulses travel to the brain through the optic nerves where edges, colors and objects are perceived.

Vision—Detailed questions

(1) What does the lens do to light?
  (a) Allows it into the eye
  (b) Blocks it from entering the eye
  (c) Focuses it
  (d) Detects shapes and colors from it
(2) What is the first layer of the eye?
  (a) Pupil
  (b) Iris
  (c) Cornea
  (d) Retina
(3) Which receptor cells enable you to see black, white, and shades of gray best?
  (a) Cones
  (b) Optic nerves
  (c) Retinal cells
  (d) Rods
(4) Which structure regulates the amount of light entering through the pupil?
  (a) Iris
  (b) Retina
  (c) Lens
  (d) Cornea
(5) What aids in seeing color and detail best?
  (a) Lens
  (b) Cones
  (c) Iris
  (d) Rods

Vision—Inference questions

(6) For most people, what is the purpose of the lens in a pair of eye glasses?
  (a) They help bend the light rays for the lens of the eyes.
  (b) They contain a certain amount of rods and cones if a retina is low in these cells.
  (c) They help the iris open and close.
  (d) They help the pupil open and close.
(7) If little light is striking the lens, what part of the eye will react to help you see better?
  (a) The cornea.
  (b) The cones.
  (c) The lens.
  (d) The iris.
(8) Which part of the eye does not work well when you're driving at night?
  (a) Lens.
  (b) Rods.
  (c) Cornea.
  (d) Cones.
(9) If you have trouble identifying a color, it may help to
  (a) Turn down the lights.
  (b) Use more rods.
  (c) Turn up the lights.
  (d) Shrink your pupil.
(10) What is the correct order of light passing through the eye?
  (a) Retina, lens, pupil, iris.
  (b) Cornea, iris, retina, pupil.
  (c) Iris, pupil, retina, cornea.
  (d) Cornea, pupil, lens, retina.

Vision test—Topic sentence cues (from Experiment 2)

Below you will see topic sentences from many of the paragraphs in the Vision text. Please attempt to recall the content from those paragraphs. You may use your own words.

A. Your eyes are made of many layers.

B. The iris is a circular structure that surrounds the pupil and can be many colors including brown, blue, grey and green.

C. After passing through the pupil, light reaches the lens.

D. After passing through the lens, the focused light rays pass through a clear, jellylike fluid in the middle of the eye. Then...

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861–876.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior, 17*, 1–12.

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (pp. 175–223). London: Academic.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*, 7–75.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*, 1118–1133.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514–527.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616.

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

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 1547–1552.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued-recall test? *Psychonomic Bulletin & Review, 13*, 826–830.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438–448.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*, 553–571.

Chi, M. T. H., deLeeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Coolidge-Stoltz, E., Cronkite, D., Graff-Haight, D., Holtzclaw, F., Jenner, J., & Cronin Jones, L. (2001). *Life science*. Upper Saddle River, NJ: Prentice Hall.

Cote, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25*, 1–53.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438–481.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research, 75*, 309–313.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*, 392–399.

Graesser, A. C., Leon, J. A., & Otero, J. (2002). Introduction to the psychology of science text comprehension. In J. Otero, J. A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 1–15). Mahwah, NJ: Lawrence Erlbaum Associates.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*, 93–103.

Hembree, R. (1988). Correlates, causes, effects and treatments of test anxiety. *Review of Educational Research, 58*, 47–77.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects with completion tests. *Memory, 19*, 290–304.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629.

Kang, S., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *The European Journal of Cognitive Psychology, 19*, 528–558.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 3–4.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17–29.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*, 227–239.

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders, 25*, 51–64.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 294–303.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*, 133–159.

Lehman, S., Schraw, G., McCrudden, M. T., & Hartley, K. (2007). Processing and recall of seductive details in scientific text. *Contemporary Educational Psychology, 32*, 569–587.

Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfeld, G. J., & Hopkins, B. L. (1971). A comparison of students studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis, 4*, 257–264.

Mayer, R. E. (2001). Introduction to multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 1–16). Cambridge University Press.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*, 516–522.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 371–385.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 954–1446.

McDaniel, M., Roediger, H. L., & McDermott, K. (2006). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review, 14*, 200–206.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*, 1–30.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43.

Millis, K., & Graesser, A. C. (1994). The time-course of constructing knowledge-based inference for scientific texts. *Journal of Memory and Language, 33*, 583–599.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., et al. (2007). *Organizing instruction and study to improve student learning (NCER 2007–2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. <http://ncer.ed.gov>.

Pellegrino, J. W. (2012). From cognitive principles to instructional practices: The devil is often in the details. *Journal of Applied Research in Memory and Cognition, 1*, 260–262.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Brighton, UK: Psychology Press.

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

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

Sanchez, C. A., & Wiley, J. (2011). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors, 51*, 730–738.

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

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Tests expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*, 264–273.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*, 252–257.

van den Broek, P., Lorch, P. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081–1087.

Voss, J. F., & Siflies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction, 14*, 45–68.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571–580.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology, 132*, 408–428.

Wiley, J., & Myers, J. L. (2003). Availability and accessibility of information and causal inferences from scientific text. *Discourse Processes, 36*, 109–129.

Wiliam, D. (2007). Keeping learning on track. In F. K. Lester, Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age Publishing.

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

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*, 995–1008.