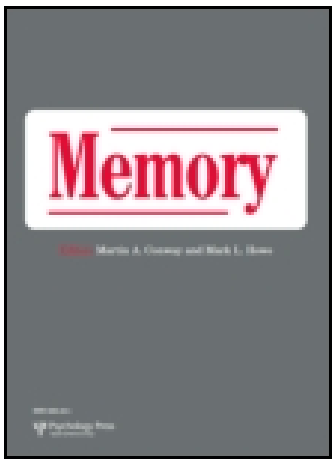

Memory

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

Retrieval practice with short-answer, multiple-choice, and hybrid tests

Megan A. Smith[a] & Jeffrey D. Karpicke[a]

[a] Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

Published online: 23 Sep 2013.

To cite this article: Megan A. Smith & Jeffrey D. Karpicke (2014) Retrieval practice with short-answer, multiple-choice, and hybrid tests, Memory, 22:7, 784-802, DOI: 10.1080/09658211.2013.831454

To link to this article: http://dx.doi.org/10.1080/09658211.2013.831454

Memory, 2014
Vol. 22, No. 7, 784–802, http://dx.doi.org/10.1080/09658211.2013.831454

Routledge
Taylor & Francis Group

Retrieval practice with short-answer, multiple-choice, and hybrid tests

Megan A. Smith and Jeffrey D. Karpicke

Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

(*Received 29 May 2013; accepted 29 July 2013*)

Retrieval practice improves meaningful learning, and the most frequent way of implementing retrieval practice in classrooms is to have students answer questions. In four experiments ($N = 372$) we investigated the effects of different question formats on learning. Students read educational texts and practised retrieval by answering short-answer, multiple-choice, or hybrid questions. In hybrid conditions students first attempted to recall answers in short-answer format, then identified answers in multiple-choice format. We measured learning 1 week later using a final assessment with two types of questions: those that could be answered by recalling information verbatim from the texts and those that required inferences. Practising retrieval in all format conditions enhanced retention, relative to a study-only control condition, on both verbatim and inference questions. However, there were little or no advantages of answering short-answer or hybrid format questions over multiple-choice questions in three experiments. In Experiment 4, when retrieval success was improved under initial short-answer conditions, there was an advantage of answering short-answer or hybrid questions over multiple-choice questions. The results challenge the simple conclusion that short-answer questions always produce the best learning, due to increased retrieval effort or difficulty, and demonstrate the importance of retrieval success for retrieval-based learning activities.

Keywords: Retrieval practice; Testing effect; Learning; Question format; Short-answer; Multiple-choice.

Practising retrieval is an effective strategy to enhance meaningful learning (e.g., Karpicke & Blunt, 2011). Retrieval practice can be implemented in the classroom through many activities, but most research to date has focused on frequent testing and quizzing in the classroom (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). If retrieval practice is to be implemented in the classroom, then it is important to know which retrieval practice formats are most effective for promoting meaningful learning. The purpose of this paper is to examine the effectiveness of various retrieval practice formats on long-term meaningful learning.

Past research has focused primarily on retrieval practice via short-answer and multiple-choice questions because these formats are frequently employed in the classroom. Multiple-choice questions require students to recognise and select a correct response among alternatives, while short-answer

Address correspondence to: Megan A. Smith, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, USA. E-mail: smith598@purdue.edu

The writing of this paper was supported in part by grants from the National Science Foundation (DUE-0941170 and DRL-1149363) and the Institute of Education Sciences in the US Department of Education (R305A110903). The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education. A portion of this project was an undergraduate honours thesis completed by Megan Smith under the instruction of Jeffrey Karpicke. We wish to thank Althea Bauernschmidt for numerous constructive discussions, Emily Boyne, Cathrine Brattain, Kait Cross, Kelli Olifirowicz, Samantha Ostler, Victor Panfil, and Nikita Saoji for help with data collection and scoring, and Philip Grimaldi for technical assistance.

questions require students to recall and produce responses. Multiple-choice and short-answer questions most readily test over verbatim or factual information, but can also be used to test higher-level concepts from Bloom's (1956) Taxonomy (Marsh, Roediger, Bjork, & Bjork, 2007). For example, they can be used for inference questions that require students to put information that they have learned together, and for application questions where students are required to take what they have learned and apply it to a new context. From an instructor's perspective, multiple-choice questions have advantages. Relative to short-answer questions, multiple-choice questions are easier to administer and grade. This is especially true with the availability of clickers and online testing systems, which can be used to administer and score multiple-choice questions in large classrooms (e.g., Mayer et al., 2009; McDaniel, Anderson, Derbish, & Morrisette, 2007). Conversely short-answer questions are more difficult to administer, and take much more time to grade.

However, some research has shown that retrieval practice using short-answer questions benefits learning more than multiple-choice questions. Short-answer questions require students to engage in more effortful and complete retrieval practice than multiple-choice questions, and more effortful retrieval has been theorised to explain why retrieval practice is effective (Pyc & Rawson, 2009). For example, Kang, McDermott, and Roediger (2007) had subjects study journal articles and then answer initial short-answer questions or answer initial multiple-choice questions. Subjects received feedback by viewing the correct answer to each question after they provided their answer. Three days later, subjects returned and took a final retention assessment in both short-answer and multiple-choice formats. On the final multiple-choice assessment, answering initial short-answer questions produced greater performance than answering initial multiple-choice questions. On a final short-answer assessment, answering initial short-answer questions produced numerically greater performance relative to answering initial multiple-choice questions, but this effect did not reach statistical significance. Importantly, the authors concluded that feedback was crucial in order to find these differences. In an experiment where feedback was not provided, practising retrieval by answering short-answer questions did not result in the best learning outcomes. The lack of a short-answer benefit is likely because initial retrieval success is frequently lower on short-answer questions than on multiple-choice questions, disadvantaging the short-answer group. Kang and his colleagues argued that feedback made up for the initial success differences between the two formats by ensuring that all subjects saw the correct answers.

Other experiments have reported a retention benefit for practising retrieval with short-answer questions relative to multiple-choice questions (e.g., Butler & Roediger, 2007; McDaniel, Anderson et al., 2007; Clariana, 2003; Duchastel, 1981; Gay, 1980; for similar findings in the adjunct questions literature see also Anderson & Biddle, 1975; Hamaker, 1986; Williams, 1965). In addition, answering multiple-choice questions exposes students to false information by presenting lures along with the correct answer, and there is some evidence that this can lead to retention of this false information (e.g., Roediger & Marsh, 2005). Results such as these have led to the conclusion that practising retrieval with short-answer questions is superior to multiple-choice questions for enhancing student learning, even if these tests are more difficult to administer.

While some have found a retention advantage for practising retrieval with short-answer questions over multiple-choice questions, the effect seems to occur only under specific circumstances. As noted above, Kang et al. (2007) reported that the difference between their short-answer and multiple-choice conditions did not reach the level of significance when the final assessment was in the short-answer format. Conversely, Gay (1980) reported the opposite result. Gay had subjects repeatedly practice retrieval with short-answer or multiple-choice questions over material in a college course and found that short-answer questions led to superior learning only when the final assessment was in short-answer format. When the final assessment was in multiple-choice format, no differences between the initial retrieval practice formats were found. It is possible to explain these results using the transfer-appropriate processing framework (Morris, Bransford, & Franks, 1977). The transfer-appropriate processing framework posits that memory will be best when the processing necessary at the time of retrieval matches the processing that was necessary at the time of encoding. The transfer-appropriate processing framework would predict that retrieval practice via short-answer tests show the greatest advantage over other formats when the final test is in short-answer format. While this explanation seems plausible for Gay's study,

the opposite was found to be true in Kang and colleagues' (2007) studies, leading to the conclusion that transfer appropriate processing cannot explain format differences.

Furthermore, there have also been studies that have failed to find a retention advantage of retrieval practice with short-answer over multiple-choice questions at all (e.g., Clariana & Lee, 2001; Duchastel & Nungester, 1982; Frase, 1968; Haynie, 1994; Williams, 1963). It is possible to explain these results; for example, Clariana and Lee (2001) and Williams (1963) found trends favouring short-answer over multiple-choice tests but their results were not statistically significant. In addition, Haynie (1994), Duchastel and Nungester (1982), and Frase (1968) did not provide feedback to their subjects. Recall that Kang and colleagues (2007) reported that feedback is necessary for retrieval practice with short-answer questions to produce more learning than multiple-choice questions. Taken together, these results cast doubt on the retrieval practice advantage of short-answer tests, or suggest that the effect only occurs under specific circumstances.

The benefit from retrieval practice might not depend on the format of the retrieval activity, but might instead depend on both retrieval difficulty and retrieval success. Some have argued that retrieval difficulty is the reason short-answer questions can produce greater retention benefits over multiple-choice questions (e.g., McDaniel, Roediger, & McDermott, 2007). The more effortful the retrieval practice, the greater the benefit. Still, success during retrieval practice is also important for final retention (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Agarwal, & Roediger, 2009). Thus if initial retrieval practice is too difficult and retrieval success is low, later retention will likely suffer. Practising retrieval with multiple-choice questions often leads to greater success relative to short-answer questions. It is possible that both retrieval difficulty and retrieval success determine whether a retrieval activity will be better in a given situation. The effects of retrieval practice will be best when retrieval is both difficult and successful (Pyc & Rawson, 2009).

One solution to these problems is to combine short-answer and multiple-choice formats, which we refer to as hybrid formats. Combining short-answer and multiple-choice formats should lead to a benefit due to the more effortful retrieval during short-answer responding, but greater success during multiple-choice responding. In educational settings these hybrid formats should have yet another benefit: the multiple-choice questions make them easier to administer and score. Park (2005) investigated the effectiveness of a hybrid testing format with sixth-grade students. He created a computerised retrieval practice system to combine the benefits of both short-answer and multiple-choice formats (see also Park & Choi, 2008). The computer first presented the question to students without alternatives so they could respond as if they were answering a short-answer question. Then, when the students were ready, multiple-choice alternatives appeared so they could find and select the answer they had already retrieved. Using this format for retrieval practice allows for quick objective scoring while still retaining the retention benefits of short-answer retrieval practice. Park found that sixth-grade students who practised retrieval via this new hybrid format performed better on a final assessment 4 days later relative to students who practised retrieval in the standard multiple-choice format. Because hybrid formats combine short-answer and multiple-choice questions they may be especially effective formats for improving meaningful learning; however, aside from Park's papers there has been very little research on this topic (for one exception see Butler, Huelser, Caruso, & Roediger, 2008).

The purpose of these experiments was to examine retrieval practice with hybrid formats, which could hold great potential for computer-based retrieval practice systems to improve meaningful learning. One of our goals was to replicate the finding that retrieval practice with a hybrid format produces more meaningful learning than a multiple-choice format, and to examine whether a hybrid format might produce more learning and retention than a short-answer format. Another goal was to see whether we could enhance the effects of hybrid formats by introducing spacing between the two retrieval attempts. In the original hybrid format (Park, 2005; Park & Choi, 2008) the questions were answered in the short-answer and then multiple-choice format immediately after one another. It is possible that spacing the repetition of questions within the hybrid format would result in greater benefits than the way hybrid formats were implemented in prior research (for a review and discussion of the spacing effect in learning and memory research see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Spacing retrieval practice has been shown to improve memory (e.g., Karpicke & Bauernsch-

midt, 2011), and spacing the same question might allow students to practise retrieval twice instead of just remembering the answer (see Jacoby, 1978). When students engage in repeated retrieval, their performance increases (e.g., Karpicke & Roediger, 2008). To meet this goal we examined two hybrid formats: a hybrid-massed format that was similar to that of Park (2005; see also Park & Choi, 2008) and a hybrid-spaced format that introduced spaced repetitions. Finally, a third goal was to measure the percentage of lures from the multiple-choice format that are produced on the final assessment in Park's method (see Roediger & Marsh, 2005).

The four experiments reported here examined the relative benefits of different retrieval practice formats on learning of meaningful educational materials. Importantly, during retrieval practice students were required to answer questions tapping conceptual knowledge that was directly stated in the text (verbatim questions) and make inferences connecting more than one concept in the text (inference questions, Experiments 1, 2, and 3; see Karpicke & Blunt, 2011). We asked if retrieval practice with short-answer questions would produce greater meaningful learning relative to multiple-choice questions, and whether a combination of the two formats would produce even greater retention. Subjects read educationally relevant texts, and then practised retrieval with one of four initial retrieval formats: a standard short-answer format, a standard multiple-choice format, or a hybrid format. Two hybrid formats were examined. The first was the *hybrid-massed* format where the short-answer and multiple-choice presentations occurred one right after the other (Park, 2005). The second was a *hybrid-spaced* format where the short-answer and multiple-choice presentations were spaced apart from one another. All subjects received feedback after practising retrieval (Kang et al., 2007). In addition we assessed the relative difficulty of our retrieval formats by recording response times to answer questions (see Benjamin, Bjork, & Schwartz, 1998; Gardiner, Craik, & Bleasdale, 1973; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007a; Pyc & Rawson, 2009). One of the reasons retrieval practice is thought to be more beneficial with short-answer questions than with multiple-choice questions is that short-answer questions are thought to induce more difficult and complete forms of retrieval (see McDaniel, Roediger, et al., 2007). We recorded response times to assess whether our short-answer format was more difficult than our multiple-choice format.

All subjects answered final short-answer questions (Experiments 1, 2, 3, and 4) or final multiple-choice questions (Experiment 4) to assess meaningful learning of the materials. The final assessment was primarily in short-answer format to require subjects to produce what they learned and not just recognise correct answers on the final assessment. However, in Experiment 4 we examined performance on both a final short-answer and multiple-choice assessment. During the final assessment students again answered verbatim questions directly tapping conceptual knowledge and inference questions requiring students to combine and integrate information. By including both direct conceptual questions and questions requiring inferences on the final assessment, we were able to ask whether different retrieval practice formats interacted with the type of knowledge being assessed, which would have important implications for educators seeking to improve meaningful learning. Experiment 4 only included verbatim questions on the final assessment, and the reasons for this will be discussed later.

EXPERIMENT 1

Method

Subjects. A total of 80 Purdue University undergraduates participated in Experiment 1. All subjects were native speakers of English.

Materials. Text materials and questions were taken from Butler, Flanagan, Roediger, and McDaniel (2007). The materials consisted of four texts, each organised into four paragraphs. An example text is provided in the Appendix. Each text covered a single topic: *Venice* (540 words), *Galileo* (534 words), *First Crusade* (590 words), and *KGB* (568 words). The order in which the texts were presented was held constant for all subjects (*Venice*, *Galileo*, *First Crusade*, and *KGB*). Each retrieval activity contained two types of questions, verbatim and inference. Example questions are provided in the Appendix. Verbatim questions had answers taken directly from the text. Inference questions had answers requiring integration of facts within a paragraph, so the answers could not be found word-for-word in the text. Eight questions were created for each text,

one verbatim and one inference question per paragraph. For the multiple-choice questions the correct answer was accompanied by four lure responses. Each response alternative was a plausible response to the question. Using each question and corresponding answer, we created one-sentence statements to use as feedback (also shown in the Appendix).

Design. A 5 (retrieval format) $\times$2 (question type) mixed factorial design was used. Retrieval format was manipulated between subjects, and 16 subjects were assigned to each condition: short-answer, multiple-choice, hybrid-massed, hybrid-spaced, and a no retrieval practice control condition. Question type (verbatim vs inference) was manipulated within subjects.

Procedure. Subjects were tested in groups of four or fewer. The experiment consisted of two sessions spaced 1 week apart. The initial session consisted of a series of study periods followed immediately by initial retrieval practice periods. During a study period, subjects studied a text on paper for 5 minutes after which the experimenter collected the text. During an initial retrieval practice period, subjects used the computer to complete eight questions corresponding to the text they just studied. The order of questions was held constant for all subjects during retrieval practice. Subjects were able to proceed to the next question by pressing the ENTER key, but the computer automatically advanced after 30 seconds. Subjects in the multiple-choice condition answered each question by selecting an answer among five alternatives and typing the corresponding number (1–5) into the computer. Subjects in the short-answer condition answered all of the questions by typing their answers into the computer. Subjects in the hybrid-massed condition first answered a question in short-answer format, and then immediately answered the same question in multiple-choice format. They continued in this manner until all eight questions were completed. This condition is quite similar to that of Park (2005); however, Park did not require that students type out their answers to the short-answer questions, and they had less time to answer the multiple-choice questions. We had our subjects type their short-answer responses so we could assess their retrieval success on the short-answer format, and we gave our subjects longer to answer the multiple-choice questions because our response alternatives were much longer than Park's one-word alternatives. Subjects in the hybrid-spaced condition answered all eight questions in short-answer format first, and then answered all eight questions in multiple-choice format. The order of the questions during the short-answer portion of the test was the same as that of the multiple-choice portion of the test. Finally, a fifth group served as a no retrieval practice control condition; these subjects did not complete an initial retrieval activity.

Response times were measured for both the multiple-choice and short-answer questions, beginning when the question was presented on the screen and ending when the subject pressed ENTER to advance to the next question. For the short-answer questions the response times include the time to read the question and type an answer. For the multiple-choice questions the response times include the time to read the question and the presented options and to select the answer. Response times to answer multiple-choice questions likely include more reading time, and response times to answer short-answer questions include more time typing out the response. We used this method in lieu of measuring the time from the initial keystroke to the completion of the answer because subjects are likely engaging in retrieval practice while reading the available responses of a multiple-choice question. By the time the subjects strike the first key during a multiple-choice question, they have already practised retrieval. Taking the full time to answer the questions seems to be the best measure of retrieval difficulty available given that the two methods of answering questions are different.

After the initial retrieval practice periods all subjects read the list of statements, which provided feedback. The no retrieval practice group read the statements after the study periods because they did not practise initial retrieval. Each statement corresponded to one question from the initial retrieval activity, and each was presented one at a time on the computer screen for 10 seconds. The statements were presented in the same order as the questions were presented. Subjects completed this procedure (study, initial retrieval practice, feedback) for four texts. After subjects completed the procedure for the fourth text they were dismissed and asked to return to the lab 1 week later for the second session.

During the second session subjects returned to the lab to take a final short-answer assessment containing all eight of the initial questions from each text. All of the questions were grouped together by text in the same fixed order from the

first session. In other words, subjects answered questions from *Venice* in the first block, *Galileo* second, *First Crusade* third, and *KGB* last. However, within each block, questions were presented in a random order determined by the computer. Each question was presented one at a time on the computer screen, and subjects were asked to type the correct answer to each question into the computer. Each question was presented for 30 seconds. After completing the final assessment, all subjects were debriefed and thanked for their participation.

Scoring. The computer scored all multiple-choice responses. One point was given when subjects selected the correct alternative and zero points were given for incorrect or no response. All short-answer data were scored by hand. One point was given for a fully correct response, half of a point for a partially correct response, and zero points for an incorrect response or no response. Different scoring procedures were used for short-answer and multiple-choice formats because it is possible to have a partially correct answer on a short-answer question, but not on a multiple-choice question. In addition, this is likely how educators would score the two formats. Most importantly, since the final assessment was all in the short-answer format, scoring of the dependent measure was held constant across all conditions. A single rater completed all scoring. As a reliability check, 20% of the data were scored a second time (across all four experiments). The correlation between the two scores was .94. All scorers were unaware of which subject produced each response and to which condition subjects belonged.

Results

All results were significant at the .05 level unless stated otherwise.

Initial performance. Table 1 shows the mean proportion correct for each initial retrieval practice format. In general, performance on the initial multiple-choice questions was similar across conditions. The multiple-choice data were entered into a 3 (retrieval format) $\times 2$ (question type) ANOVA with repeated measures on the second factor. There were no main effects of retrieval format or question type and there was no interaction (all Fs < 1). Performance on the initial short-answer questions was similar across conditions as well. A 3 (retrieval format) $\times 2$ (question type) ANOVA with repeated measures on the second factor was also performed on the short-answer data. There was no main effect of retrieval format ($F < 1$). However, there was a main effect of question type, $F(1, 45) = 27.66$, $\eta_p^2 = .38$. On the initial short-answer format, performance was higher for the verbatim questions ($M = .45$) than for the inference questions ($M = .33$). There was no interaction, $F(2, 45) = 2.88$, $p = .07$. We also compared initial performance between the short-answer and multiple-choice groups, and found that the multiple-choice group performed better than the short-answer group on the initial tests, $F(1, 30) = 58.50$, $\eta_p^2 = .66$. However, we provided feedback, which should help ameliorate the initial success disadvantage for the short-answer group (Kang et al., 2007).

Table 2 shows mean response times to answer questions for each retrieval format. Only response times from questions answered correctly are reported (see Karpicke & Roediger, 2007a; Pyc & Rawson, 2009). We directly compared response times on the multiple-choice and short-answer formats to ensure that retrieval during our short-answer questions was in fact more difficult than retrieval during our multiple-choice questions. A 2 (retrieval format: multiple-choice vs short-answer) $\times 2$ (question type) ANOVA revealed that response times to answer multiple-choice questions ($M = 11.3$ seconds) were faster than

TABLE 1
Mean proportion correct on initial retrieval activities and the final assessment in Experiment 1

	Initial retrieval practice		*Final assessment*
Condition and question type	*Short-answer*	*Multiple-choice*	*Short-answer*
Verbatim questions			
No retrieval practice	–	–	.23 (.03)
Multiple-choice	–	.81 (.05)	.52 (.06)
Short-answer	.39 (.05)	–	.42 (.04)
Hybrid-massed	.46 (.04)	.78 (.03)	.53 (.05)
Hybrid-spaced	.49 (.06)	.81 (.04)	.52 (.07)
Inference questions			
No retrieval practice	–	–	.25 (.04)
Multiple-choice	–	.79 (.04)	.44 (.04)
Short-answer	.35 (.04)	–	.41 (.04)
Hybrid-massed	.29 (.04)	.80 (.03)	.46 (.05)
Hybrid-spaced	.36 (.05)	.77 (.05)	.46 (.06)

Standard errors in parentheses.

TABLE 2
Mean response times for correctly answered questions during initial retrieval for Experiment 1

	Initial retrieval practice	
Condition and question type	*Short-answer*	*Multiple-choice*
Verbatim questions		
Multiple-choice	–	8.3 (0.5)
Short-answer	10.3 (0.7)	–
Hybrid-massed	9.5 (0.7)	5.6 (0.4)
Hybrid-spaced	10.0 (0.7)	7.1 (0.3)
Inference questions		
Multiple-choice	–	14.2 (0.5)
Short-answer	18.1 (1.1)	–
Hybrid-massed	18.0 (0.8)	10.7 (0.8)
Hybrid-spaced	18.3 (1.0)	12.2 (0.6)

Response times in seconds. Standard errors in parentheses.

response times to answer short-answer questions ($M = 14.3$ seconds); $F(1, 30) = 10.74$, $\eta_p^2 = .26$. This is consistent with the idea that retrieval during short-answer questions involves more effort than multiple-choice questions. The response times to answer verbatim questions ($M = 9.3$ seconds) were faster than to answer inference questions ($M = 16.1$ seconds); $F(1, 30) = 151.92$, $\eta_p^2 = .84$. There was no interaction, $F(1, 30) = 2.84$, $p = .10$.

Final performance. The far right column of Table 1 shows the mean proportion correct on the final short-answer assessment. Practising retrieval improved performance on the final assessment. Subjects in all retrieval practice conditions performed better than those in the no retrieval practice condition on the final assessment when measured with both verbatim and inference questions; all Fs$(1, 30) > 8.95$, ps $< .01$. A 4 (retrieval format) $\times 2$ (question type) ANOVA was performed on the four groups that practised retrieval. There was no main effect of retrieval format ($F < 1$); the format of initial retrieval practice did not have an effect on meaningful learning. There was a main effect of question type, $F(1, 60) = 11.76$, $\eta_p^2 = .16$, indicating that performance was higher for verbatim questions ($M = .50$) than for inference questions ($M = .44$). There was no interaction ($F < 1$).

Lure intrusions on the final assessment. We examined the final assessment for lure intrusions to see if any of our retrieval practice formats led subjects to produce false information (Roediger & Marsh, 2005). For each incorrect question on the final assessment we recorded whether the response provided by the subject was originally a lure from the initial multiple-choice question. Subjects in the multiple-choice and hybrid-massed conditions produced lures 17% of the time, and those in the hybrid-spaced condition produced lures 14% of the time. However, even though subjects in the no retrieval practice and short-answer conditions never saw the lures, they still produced them 13% and 8% of the time respectively. This might have occurred because many of the lures were plausible responses that might have been produced, by chance, after reading the passage. These data were submitted to a one-way ANOVA, and the analysis indicated there were no differences among conditions, $F(4, 75) = 1.96$, $p = .11$.

Discussion

Experiment 1 demonstrated that when subjects practised retrieval after studying they performed better on a final assessment 1 week later relative to subjects who did not practise retrieval. This was true for both verbatim questions that tapped conceptual knowledge and inference questions that required subjects to integrate information from the studied passage. Importantly, the initial retrieval practice format did not seem to matter for meaningful learning. Practising retrieval with short-answer questions did not lead to greater learning than practising retrieval with multiple-choice questions even though retrieval with short-answer questions was more difficult than retrieval with multiple-choice questions as assessed by response times during retrieval practice, so difficulty was not related to subsequent retention (see also Karpicke & Bauernschmidt, 2011). Further, there was no advantage of the hybrid formats over the other formats even though students in the hybrid conditions answered each question twice while those in the short-answer and multiple-choice conditions only answered each question once. It is likely that, even in the hybrid-spaced condition, the questions were not spaced enough to induce repeated retrieval, but instead subjects simply remembered the answer (see Jacoby, 1978). It is noteworthy that practising retrieval with multiple-choice questions produced a substantial learning advantage. Retrieval practice with multiple-choice questions did not lead to a negative suggestion effect on the final test, likely because direct feedback was provided. That is, students in our experiment did not learn false information from our multiple-choice tests. Multi-

ple-choice tests have had a bad reputation, but our results indicate that multiple-choice tests can be just as effective as other forms (see also Little, Bjork, Bjork, & Angello, 2012).

EXPERIMENT 2

In Experiment 1 practising retrieval improved meaningful learning, but we did not observe differences in the effectiveness of initial retrieval formats (short-answer, multiple-choice, or hybrid formats). Although other authors have reported finding no differences between retrieval formats on learning (e.g., Duchastel & Nungester, 1982), the results are still surprising in light of recent statements about inherent advantages of retrieval practice via short-answer questions (e.g., McDaniel, Roediger, et al., 2007). Therefore we sought to repeat Experiment 1 with a few slight modifications to help bring out potential mnemonic differences among initial retrieval formats. We reasoned that subjects might have been encumbered by the task of reading four texts (totalling over 2000 words), with several concepts per text, and then answering fairly demanding inference questions. In Experiment 2 subjects read two texts and were given more time to study for each text to help remove some of the demand placed on subjects and to ensure that our results were not due to a lack of study time.

Method

Subjects. A total of 100 Purdue University undergraduates participated in Experiment 2. All subjects were native speakers of English, and none had participated in Experiment 1.

Materials. Materials included two of the four texts from Experiment 1: *Venice* and *The First Crusade*. The same questions and feedback statements from these texts from Experiment 1 were used.

Design and procedure. The design was the same as in Experiment 1, and 20 subjects were assigned to each retrieval format condition. Subjects again proceeded through a series of study, retrieval practice, and feedback periods, but this time they only repeated the procedure for two texts. During study periods subjects were instructed to study each text for 10 minutes. During the final assessment each question was presented for 45 seconds to ensure that subjects had enough time to respond to each question. Otherwise the procedure was the same as in Experiment 1.

Results

Initial performance. Table 3 shows the mean proportion correct for each initial retrieval practice format. The initial multiple-choice data were entered into a 3 (retrieval format) $\times 2$ (question type) ANOVA with repeated measures on the second factor. There was a marginal main effect of retrieval format, $F(2, 57) = 2.87$, $p = .07$, and a main effect of question type, $F(1, 57) = 11.68$, $\eta_p^2 = .17$; performance was higher for the verbatim questions ($M = .84$) than for the inference questions ($M = .78$) on the initial multiple-choice questions. There was no interaction ($F < 1$). A 3 (retrieval format) $\times 2$ (question type) ANOVA with repeated measures on the second factor was also performed on the short-answer data. There was a main effect of retrieval format, $F(2, 57) = 3.57$, $\eta_p^2 = .11$, indicating that there were differences in performance on the initial short-answer questions across retrieval format conditions. Subjects in the short-answer condition ($M = .48$) scored higher on the initial short-answer questions than those in the hybrid-massed condition ($M = .34$); $F(1, 38) = 5.84$, $\eta_p^2 = .13$, and the hybrid-spaced condition ($M = .36$); $F(1, 38) =$

TABLE 3
Mean proportion correct on initial retrieval activities and the final assessment in Experiment 2

	Initial retrieval practice		*Final assessment*
Condition and question type	*Short-answer*	*Multiple-choice*	*Short-answer*
Verbatim questions			
No retrieval practice	–	–	.23 (.04)
Multiple-choice	–	.90 (.02)	.49 (.05)
Short-answer	.58 (.05)	–	.54 (.04)
Hybrid-massed	.43 (.05)	.80 (.04)	.42 (.05)
Hybrid-spaced	.43 (.04)	.82 (.04)	.45 (.05)
Inference questions			
No retrieval practice	–	–	.22 (.04)
Multiple-choice	–	.83 (.02)	.49 (.03)
Short-answer	.37 (.04)	–	.47 (.05)
Hybrid-massed	.26 (.03)	.73 (.04)	.36 (.05)
Hybrid-spaced	.28 (.04)	.78 (.03)	.46 (.05)

Standard errors in parentheses.

4.61, $\eta_p^2 = .11$. There was also a main effect of question type, $F(1, 57) = 56.96$, $\eta_p^2 = .50$; performance was again higher for the verbatim questions ($M = .48$) than for the inference questions ($M = .30$). There was no interaction ($F < 1$). Again we compared initial performance between the short-answer and multiple-choice groups, and found that the multiple-choice group performed better than the short-answer group on the initial tests, $F(1, 30) = 81.15$, $\eta_p^2 = .68$.

We analysed response times to answer questions for each retrieval format in the same way as in Experiment 1 using correct responses only, and these values are shown in Table 4. Response times from the multiple-choice and short-answer groups were analysed using a 2 (retrieval format) $\times$ 2 (question type) ANOVA. Response times to answer multiple-choice questions ($M = 12.3$ seconds) were faster than response times to answer short-answer questions ($M = 14.9$ seconds); $F(1, 37) = 7.90$, $\eta_p^2 = .18$. This is again consistent with the idea that retrieval during short-answer questions involves more effort than retrieval during multiple-choice questions. There was also a main effect of question type, $F(1, 37) = 173.66$, $\eta_p^2 = .82$; once again the response times to answer verbatim questions ($M = 10.1$ seconds) were faster than to answer inference questions ($M = 17.1$ seconds). There was no interaction ($F < 1$).

Final performance. The far right column of Table 3 shows the mean proportion correct on the final short-answer assessment. As in Experiment 1, practising retrieval improved performance on the final assessment. Subjects in all retrieval practice conditions performed better than the no retrieval practice group for both types of questions; all Fs > 5.06, ps $< .03$. A 4 (retrieval format) $\times$ 2 (question type) ANOVA was performed on the four groups that practised retrieval. There was no main effect of retrieval format, $F(3, 76) = 1.59$, $p = .20$, or question type, $F(1, 76) = 1.88$, $p = .17$, and no interaction, $F(1, 76) = 1.08$, $p = .36$.

TABLE 4
Mean response times for correctly answered questions during initial retrieval for Experiment 2.

	Initial retrieval practice	
Condition and question type	*Short-answer*	*Multiple-choice*
Verbatim questions		
Multiple-choice	–	9.0 (0.4)
Short-answer	11.3 (0.8)	–
Hybrid-massed	12.0 (0.7)	6.1 (0.3)
Hybrid-spaced	10.2 (0.6)	7.3 (0.5)
Inference questions		
Multiple-choice	–	15.7 (0.6)
Short-answer	18.5 (1.0)	–
Hybrid-massed	18.5 (1.0)	10.8 (0.6)
Hybrid-spaced	18.2 (0.9)	13.1 (0.6)

Response times in seconds. Standard errors in parentheses.

Lure intrusions on the final assessment. Again, we examined the number of lures produced on the final assessment. For questions answered incorrectly, subjects in the multiple-choice, hybrid-massed, and hybrid-spaced conditions produced lures 11%, 15%, and 15% of the time, respectively. However, subjects in the no retrieval practice and short-answer conditions produced lures 13% and 11% of the time. A one-way ANOVA indicated that there were no differences among the conditions ($F < 1$). Thus retrieval practice with multiple-choice questions did not cause subjects to produce false information.

Discussion

Experiment 2 replicated the results of Experiment 1. Again practising retrieval after studying produced greater performance on a final short-answer assessment 1 week later compared to subjects who did not practise retrieval. Importantly, the format of the initial retrieval activity did not matter for learning even though we observed differences in retrieval difficulty between the initial short-answer and multiple-choice questions as measured by initial response times. An initial multiple-choice test still produced a retrieval practice benefit as large as that produced by an initial short-answer test without producing negative suggestion effects.

ANALYSIS ACROSS EXPERIMENTS 1 AND 2

Given that the two experiments were very similar and both failed to reveal differences among the retrieval practice conditions, we conducted an analysis combining the results of the two experiments to see if an effect would emerge with a more powerful analysis. Because the same essential design was used for both Experiment 1 and Experiment 2, we combined the final assessment results from the two experiments. The combined analysis consequently included a total of 144

TABLE 5
Mean proportion correct on the final assessments, collapsed across Experiment 1 and Experiment 2

	Verbatim questions	*Inference questions*
No retrieval practice	.23 (.03)	.23 (.03)
Multiple-choice	.50 (.04)	.47 (.03)
Short-answer	.49 (.03)	.44 (.03)
Hybrid-massed	.47 (.03)	.40 (.03)
Hybrid-spaced	.48 (.04)	.46 (.04)

Standard errors are in parentheses.

subjects, with 36 subjects in each of the four retrieval practice conditions. Mean proportions correct on the final short-answer assessment for each condition are shown in Table 5.

A 4 (retrieval format) $\times 2$ (question type) ANOVA showed that there were no differences among the four retrieval practice conditions ($F < 1$). There was a main effect of question type, $F(1, 140) = 7.94$, $\eta_p^2 = .05$; overall, the proportion correct was greater for verbatim questions ($M = .48$) than it was for inference questions ($M = .44$). There was no interaction ($F < 1$). Even with 36 subjects in each condition there were still no differences in final performance.

EXPERIMENT 3

In the first two experiments we did not observe differences in meaningful learning among short-answer, multiple-choice, or hybrid retrieval practice formats. However, in both experiments retrieval success was greater in multiple-choice conditions than in short-answer conditions. Differences between short-answer and multiple-choice retrieval activities are often explained by difficulty. The response time analyses on our initial data suggested that retrieval during our short-answer questions was more difficult than during our multiple-choice questions. However, performance on the multiple-choice questions was generally much higher than performance on the initial short-answer questions. It is possible that the differences in initial success caused the lack of format effects even though feedback was provided because both difficulty and retrieval success are important for retrieval practice to be maximally effective. So, the purpose of Experiment 3 was to increase retrieval success in the short-answer condition to see if the advantage of short-answer retrieval practice would occur under these conditions. Subjects studied a text, practised retrieval, restudied the text, and then practised retrieval a second time. This procedure provided subjects with the opportunity to restudy the material (as feedback) and allowed subjects to reattempt retrieval practice. Then, 1 week later subjects returned for the final short-answer assessment.

Method

Subjects and materials. A total of 48 Purdue University undergraduates participated in Experiment 3. All subjects were native speakers of English, and none had participated in Experiment 1 or Experiment 2. The materials were the same as in Experiment 2 (*Venice* and *The First Crusade*).

Design. A 2 (retrieval format) $\times 2$ (question type) mixed factorial design was used. A total of 24 students were assigned to one of two retrieval format conditions: short-answer or multiple-choice. Question type was manipulated within subjects.

Procedure. In the initial session subjects studied the text, practised retrieval, restudied the text, and then practised retrieval a second time. Half of the subjects answered short-answer questions and half answered multiple-choice questions during retrieval practice. During study and restudy periods subjects were instructed to study the text on the computer for 5 minutes. Students answered questions in the same way as in the first two experiments, except that they clicked a ''next'' button to move on after each question. The computer automatically advanced after 60 seconds. Unfortunately response times were not collected for Experiment 3. Subjects completed this procedure (study, retrieval practice, restudy, retrieval practice) for two texts, and then were dismissed from the initial session. Subjects returned for the final short-answer assessment 1 week later. The final assessment was the same as in Experiment 2.

Results

Initial performance. Table 6 shows initial performance for each retrieval format. The initial multiple-choice data were entered into a 2 (retrieval period) $\times 2$ (question type) ANOVA with repeated measures on both factors. There

was a main effect of retrieval period, $F(1, 23) = 26.91$, $\eta_p^2 = .54$; performance was higher on the second retrieval activity ($M = .92$) than the first retrieval activity ($M = .78$). There was also a main effect of question type, $F(1, 23) = 13.11$, $\eta_p^2 = .36$; performance was higher on verbatim questions ($M = .89$) than inference questions ($M = .81$). There was no interaction ($F < 1$). A 2 (retrieval period) $\times$2 (question type) ANOVA with repeated measures on both factors was also performed on the short-answer data. There was a main effect of retrieval period, $F(1, 23) = 102.46$, $\eta_p^2 = .82$; performance was higher on the second retrieval activity ($M = .65$) than the first retrieval activity ($M = .41$). There was also a main effect of question type, $F(1, 23) = 74.47$, $\eta_p^2 = .76$; performance was higher on verbatim questions ($M = .66$) than inference questions ($M = .40$). There was only a marginal interaction, $F(1, 23) = 3.19$, $p = .09$. Importantly, by providing subjects with the opportunity to re-study the text and practice retrieval again, we were able to increase their retrieval success.

We also compared short-answer and multiple-choice performance on the second retrieval period. The multiple-choice group still outperformed the short-answer group on initial verbatim questions, $F(1, 46) = 36.17$, $\eta_p^2 = .44$. However, the performance gap between the two formats was 13%, down from 42% and 32% in Experiments 1 and 2 respectively. The multiple-choice group still outperformed the short-answer group on the inference questions as well, $F(1, 46) = 85.64$, $\eta_p^2 = .65$, and the difference between the two groups was about the same as it was in the previous experiments (37% in Experiment 3, compared to 44% and 46% in Experiments 1 and 2 respectively). Although the procedures used in Experiment 3 did not successfully equate initial retrieval success, at least for the verbatim questions the gap was closed a great deal.

TABLE 6
Mean proportion correct on initial retrieval activities and the final assessment in Experiment 3

	Initial retrieval practice		*Final assessment*
Condition and question type	*Period 1*	*Period 2*	*Short-answer*
Verbatim questions			
Multiple-choice	.82 (.05)	.97 (.02)	.56 (.05)
Short-answer	.52 (.04)	.80 (.02)	.64 (.04)
Inference questions			
Multiple-choice	.75 (.04)	.87 (.03)	.51 (.04)
Short-answer	.30 (.03)	.50 (.03)	.40 (.04)

Standard errors in parentheses.

Final performance. Table 6 also shows the mean correct on the final assessment for each retrieval format. A 2 (retrieval format) $\times$2 (question type) mixed factorial ANOVA with repeated measures on the second factor was performed on the final performance data. This analysis revealed a main effect of question type, $F(1, 46) = 36.37$, $\eta_p^2 = .44$, because performance on the verbatim questions ($M = .60$) was higher than performance on the inference questions ($M = .45$). There was no overall main effect of retrieval format ($F < 1$). However, these results were qualified by an interaction, $F(1, 46) = 13.86$, $\eta_p^2 = .23$. For the verbatim questions there was a slight advantage of answering initial short-answer questions over answering initial multiple-choice questions, but this advantage was small and did not reach significance, $F(1, 46) = 1.34$, $p = .25$. However, a different pattern of results was found for the inference questions. For these questions, subjects in the multiple-choice condition performed better on the final assessment than subjects in the short-answer condition, $F(1, 46) = 3.97$, $p = .05$, $\eta_p^2 = .08$. Subjects achieved greater success during the initial retrieval activity, but performance on the multiple-choice questions was still greater than on the short-answer questions. When the questions were more difficult (i.e., the inference questions) the greater success allowed subjects in the multiple-choice condition to outperform those in the short-answer condition on the final assessment, an effect in the opposite direction of the results sometimes reported in the literature (e.g., Kang et al., 2007).

Lure intrusions on the final assessment. Again we examined the number of lures produced on the final assessment. When a question was answered incorrectly, subjects in the multiple-choice condition produced lures 22% of the time while subjects in the short-answer condition produced lures only 8% of the time. Subjects who answered initial multiple-choice questions produced more incorrect information previously presented to them relative to those who answered short-answer questions, $F(1, 46) = 9.28$; $\eta_p^2 = .17$. Retrieval practice with multiple-choice questions caused a negative suggestion effect in this experiment.

Discussion

The purpose of Experiment 3 was to increase retrieval success in the short-answer condition to attempt to bring out learning differences between short-answer and multiple-choice formats. We attempted to increase retrieval success by providing students with the opportunity to restudy the full text after their first retrieval practice attempt, and then providing a second opportunity for students to practise retrieval. Looking at the second retrieval practice attempt, this procedure closed the gap between short-answer and multiple-choice performance compared to the first two experiments, at least for the verbatim questions. Whereas performance differences were 42% and 32% between short-answer and multiple-choice formats in Experiments 1 and 2 respectively, the difference between the two formats in Experiment 3 was 13% by the second retrieval practice period. On the final assessment the short-answer group performed 8% better than the multiple-choice group, though this difference was not statistically significant. For inference questions, the performance difference was still high (37% in Experiment 3, compared to 44% and 46% in Experiments 1 and 2, respectively). Thus it appears that doing more to equate the success differences between test formats may lead to differences on a later assessment. We also found a negative suggestion effect in Experiment 3. Those in the multiple-choice condition produced more lures on the final assessment than those in the short-answer condition. The negative suggestion effect may have occurred in Experiment 3 because direct feedback was not provided; instead, subjects were given a restudy opportunity. Feedback is important for ameliorating the negative effects of multiple-choice tests (Butler et al., 2006). These data suggest that a restudy opportunity is not enough to reduce negative suggestion effects.

EXPERIMENT 4

In three experiments we have shown that using tests to practise retrieval improves learning, but the format of a test does not much matter for these effects. Attempts to close the initial gap between short-answer and multiple-choice success revealed a small advantage in favour of the short-answer format. Therefore the purpose of Experiment 4 was to try again to equate initial performance on the short-answer and multiple-choice tests as closely as possible. To do this we used a procedure similar to that of Experiment 1. Subjects read a text and then answered questions in one of three test formats: multiple-choice, short-answer, or hybrid. A fourth no retrieval practice group was also included. For this Experiment we used four texts that were shorter than the ones used in the previous experiments to try to make it easier for the subjects to perform well on the initial tests. We created a number of questions to go along with each text and measured success on these questions in both a short-answer and multiple-choice format in a pilot. Questions used for Experiment 4 resulted in performance that was as close to equal between the two formats as possible. During the pilot the multiple-choice group successfully answered 89% and the short-answer group successfully answered 83% correct ($F < 1$) for the 32 questions that we ultimately used in Experiment 4. In addition we used only verbatim questions because the data from Experiment 3 suggested that it is easier to equate performance across the two formats with verbatim questions than with inference questions. As in Experiments 1 and 2, direct feedback was provided to all subjects in the form of the correct answer embedded within a sentence. On a final assessment 1 week later, subjects completed some questions in a short-answer format and others in a multiple-choice format.

Method

Subjects. Subjects were 144 Purdue University undergraduates. None had participated in the previous experiments reported here.

Materials. Four text materials were taken from the reading comprehension section of a test-preparation book for the Test of English as a Foreign Language (TOEFL; Roediger & Karpicke, 2006; Rogers, 2001). An example text is provided in the Appendix. Each text covered a single topic: *The Sun* (256 words), *Sea Otters* (275 words), *Early History of Jazz* (219 words), and *Bessemer Steel* (270 words). The order in which the texts were presented was held constant for all subjects (*The Sun*, *Sea Otters*, *Early History of Jazz*, and *Bessemer Steel*).

Prior to the experiment, 11 questions were written in multiple-choice and short-answer

format for piloting. For the multiple-choice questions the correct answer was accompanied by four lure responses. The lure responses were carefully constructed such that each one was a plausible answer to the question (see Little et al., 2012). Then 12 subjects, none of whom participated in Experiment 4 or any previously reported experiments, completed the same initial study and testing phase that was described in Experiment 1. Eight questions were then selected for each text based on the pilot results. The questions selected best matched initial success during testing for the multiple-choice and short-answer conditions (i.e., the three questions from each text that had the largest differences in performance between short-answer and multiple-choice formats were not used in the experiment). These eight questions were used for Experiment 4. Example questions are provided in the Appendix. Using each question and corresponding correct answer, we created one-sentence statements to use as feedback (also shown in the Appendix).

Design. A 4 (retrieval format) $\times 2$ (final test format) mixed factorial design was used. A total of 36 students were assigned to each of the four retrieval format conditions: short-answer, multiple-choice, hybrid, and a no retrieval practice control condition. The hybrid condition was the same as the hybrid-massed condition from Experiments 1 and 2. Final test format (short-answer vs multiple-choice) was manipulated within subjects, with two texts assigned to the multiple-choice format and two to the short-answer format. The texts assigned to the two final test formats were fully counterbalanced across subjects.

Procedure. The procedure was very similar to that of Experiment 1, with a few modifications to each session listed below. First, subjects studied the text on the computer. Second, during the initial retrieval practice period, subjects were instructed to click the next button on the screen when they were done answering each question. The next button appeared on the screen after 1 second to help ensure that students did not continuously click next through the experiment. In this experiment we did not enforce a maximum time limit on responding.

During the second session we assessed final performance using a short-answer and multiple-choice test. For each subject two texts were tested in the short-answer format and two in the multiple-choice format. Subjects were instructed to click the next button on the screen when they had finished answering each question. The next button appeared on the screen after 4 seconds to help ensure that students began to answer each question. Again, no maximum time limit was enforced.

Results

Initial performance. Table 7 shows the mean proportion correct for each initial retrieval practice format. Performance on the initial multiple-choice questions was extremely similar across conditions. The multiple-choice data were entered into a one-way ANOVA and showed that there was no difference between the multiple-choice and hybrid conditions on the multiple-choice questions ($F < 1$). Performance on the initial short-answer questions was extremely similar across conditions as well. A one-way ANOVA revealed that there was no difference between the short-answer and hybrid conditions on short-answer questions ($F < 1$). We also compared initial performance in the short-answer and multiple-choice conditions. A one-way ANOVA revealed that performance was higher in the multiple-choice condition ($M = .83$) than in the short-answer condition ($M = .72$); $F(1, 70) = 8.81$, $\eta_p^2 = .11$. One of the purposes of Experiment 4 was to match performance on the initial short-answer and multiple-choice questions. While performance was not perfectly matched, the difference in performance was much closer in this experiment relative to the others reported in this paper. The difference between the two conditions was only 10.8% in this experiment, whereas in the first two experiments the difference between the

TABLE 7
Mean proportion correct on initial retrieval activities and final assessments in Experiment 4

	Initial retrieval practice		*Final assessment*	
Condition	*Short-answer*	*Multiple-choice*	*Short-answer*	*Multiple-choice*
No retrieval practice	–	–	.41 (.03)	.60 (.03)
Multiple-choice	–	.83 (.02)	.50 (.03)	.76 (.03)
Short-answer	.72 (.03)	–	.59 (.04)	.80 (.03)
Hybrid	.70 (.03)	.83 (.02)	.63 (.04)	.80 (.03)

Standard errors in parentheses.

short-answer and multiple-choice conditions was around 30–40%.

We analysed response times to answer questions for each retrieval format in the same way as in Experiments 1 and 2 using correct responses only, and these values are shown in Table 8. As in Experiments 1 and 2, correct response times during the short-answer initial test ($M = 13.5$ seconds) were longer than correct response times during the multiple-choice format ($M = 8.8$ seconds); $F(1, 70) = 12.05$, $\eta_p^2 = .15$.

Final performance. The two right-hand columns of Table 7 show the mean proportion correct on the final test in both multiple-choice and short-answer formats. Subjects in all retrieval practice conditions performed better than those in the no retrieval practice condition on the final short-answer and multiple-choice assessments; all Fs$(1, 70) > 4.04$, ps $< .05$. A one-way ANOVA performed on the three conditions that practiced retrieval revealed that, on the final multiple-choice assessment, there was no main effect of retrieval format ($F < 1$). However, a different pattern of results was found for the short-answer assessment. On the short-answer assessment there was a main effect of retrieval format, $F(2, 105) = 3.34$, $\eta_p^2 = .06$. Least significant difference post-hoc comparisons revealed that students in the hybrid condition ($M = .63$) performed better than those in the multiple-choice condition ($M = .50$), $F(1, 70) = 6.59$, $\eta_p^2 = .09$, and those in the short-answer condition ($M = .59$) performed marginally better than those in the multiple-choice condition, $F(1, 70) = 3.34$, $p = .07$, $\eta_p^2 = .05$. No other pairwise comparisons reached significance. Thus we have replicated the results from Park (2005; Park & Choi, 2008): practising retrieval with a hybrid test format improves learning relative to taking a multiple-choice test. In addition we found a marginal advantage of practising retrieval using a short-answer test over using a multiple-choice test on later performance when performance was more closely equated during retrieval practice and exact feedback was provided.

TABLE 8
Mean response times for correctly answered questions during initial retrieval for Experiment 4

	Initial retrieval practice	
Condition and question type	*Short-answer*	*Multiple-choice*
Multiple-choice	–	8.8 (0.6)
Short-answer	13.5 (1.2)	–
Hybrid	11.8 (0.6)	5.0 (0.3)

Response times in seconds. Standard errors in parentheses.

Lure intrusions on the final assessment. Again, we analysed the proportion of lures produced on the final short-answer test. Four subjects were not included in the analysis because they answered all of the questions on the final short-answer test correctly. For questions answered incorrectly, subjects in the multiple-choice and hybrid conditions produced lures 28% and 29% of the time, respectively. However, subjects in the no retrieval practice and short-answer conditions produced lures 22% and 19% of the time. A one-way ANOVA indicated that there were no differences among the conditions, $F(3,136) = 1.78$, $p = .15$. As in Experiments 1 and 2, when exact feedback was provided after the initial retrieval practice test, retrieval practice with multiple-choice questions did not lead subjects to produce false information.

GENERAL DISCUSSION

The purpose of these experiments was to investigate the relative benefits of practising retrieval with various retrieval formats on meaningful learning. In addition we sought to examine how retrieval practice affects performance on two types of questions: verbatim questions directly tapping conceptual knowledge and inference questions requiring subjects to combine information. Practising retrieval improved performance on both verbatim and inference questions on a short-answer assessment 1 week later (Experiments 1 and 2). In fact, practising retrieval improved performance by nearly double over the no retrieval practice group in the first two experiments. Even practising retrieval with multiple-choice questions doubled later performance relative the no-test condition, demonstrating that practising retrieval with multiple-choice tests can be a powerful way to improve learning. In Experiment 4 we assessed learning using both a final short-answer and a final multiple-choice test, and showed that practising retrieval improved performance on both final assessments. Table 9 shows the effect sizes (d) for all independent comparisons of the retrieval practice formats relative to the no retrieval practice control conditions. The bottom row shows the overall effect size, calculated using weighted effect sizes and a random effects meta-analysis model (Cummings,

TABLE 9
Effect size *d* for each retrieval practice condition compared to the no retrieval practice control for Experiments 1, 2, 4, and overall

	Multiple-choice	*Short-answer*	*Hybrid-massed*	*Hybrid-spaced*
Experiment 1	1.41 [0.69, 2.13]	1.31 [0.59, 2.03]	1.61 [0.89, 2.33]	1.22 [0.50, 1.94]
Experiment 2	1.73 [1.09, 2.37]	1.72 [1.08, 2.36]	1.00 [0.36, 1.64]	1.36 [0.72, 2.00]
Experiment 4	0.82 [0.35, 1.29]	1.15 [0.68, 1.62]	1.24 [0.77, 1.71]	–
Overall	**1.24 [0.68, 1.79]**	**1.31 [0.95, 1.66]**	**1.22 [0.87, 1.58]**	**1.27 [0.77, 1.76]**

Effects are collapsed across all within-subjects variables: question type (Experiments 1 and 2) and final assessment format (Experiment 4). 95% confidence intervals around *d* are in brackets.

2012; see also Smith, Roediger, & Karpicke, 2013). The overall effects of practising retrieval in any format on later performance were large. Further, the effects of retrieval practice among the different formats were all very similar to one another. Even though multiple-choice tests are sometimes thought to produce little to no retrieval practice effects, our data suggest that retrieval practice with multiple-choice tests can be quite effective at improving later performance. Recent research by Little and colleagues (2012) provides converging evidence. In addition, in educational settings questions from the practice tests do not often appear on the final assessment tests; however, Little and colleagues have shown that multiple-choice tests can produce learning benefits even when the assessment questions are different from those on the initial retrieval-practice test.

However, across four experiments we found that initial retrieval practice format only mattered for learning under very specific circumstances. Specifically, learning differences between formats only emerged when initial retrieval practice success was as similar as possible, and when direct feedback was provided (as opposed to providing additional opportunities to restudy the material after retrieval practice). In the first three experiments we found that the initial retrieval practice format did not much matter for learning. In a fourth experiment we more closely equated initial retrieval success and provided direct feedback in the form of the correct answer. In this experiment we found an advantage of the hybrid format over the multiple-choice format, and a marginal advantage of the short-answer format over the multiple-choice format when a final short-answer assessment was used. Table 10 shows the effect sizes (*d*) for independent comparisons between retrieval formats, and again the bottom row shows the overall effect across the experiments. Contrary to the retrieval practice effects from Table 9, the overall effect sizes between retrieval formats were all very close to zero.

One explanation for the differences sometimes found between retrieval practice formats is the difficulty of the retrieval attempt afforded by the retrieval activity. Some have argued that retrieval practice produces desirable difficulties that enhance long-term retention (Bjork, 1994; 1999; Pyc & Rawson, 2009). Short-answer questions are thought to be more difficult than multiple-choice questions, and thus could produce greater benefits. However, success of retrieval is also important for later performance (e.g., Butler et al., 2006; Marsh et al., 2009). Our response time data indicated that our short-answer questions were more difficult than our multiple-choice questions. However, most of the time, subjects performed better on the multiple-choice questions than they did on the short-answer questions. We do not doubt that retrieval practice with short-answer questions requires more effortful retrieval and

TABLE 10
Effect size *d* for the retrieval practice conditions compared to one another for all four experiments and overall

	SA – MC	*Hybrid-M – SA*	*Hybrid-S – SA*	*Hybrid-M – MC*	*Hybrid-S – MC*
Experiment 1	−0.37 [−1.09, 0.35]	0.47 [−0.25, 1.19]	0.34 [−0.38, 1.06]	0.06 [−0.66, 0.78]	0.03 [−0.69, 0.75]
Experiment 2	0.08 [−0.56, 0.72]	−0.63 [−1.27, 0.01]	−0.26 [−0.90, 0.38]	−0.58 [−1.22, 0.06]	−0.19 [−0.83, 0.45]
Experiment 3	−0.07 [−0.65, 0.51]	–	–	–	–
Experiment 4	0.40 [−0.07, 0.87]	0.12 [−0.35, 0.59]	–	0.52 [0.05, 0.99]	–
Overall	**0.07 [−0.24, 0.38]**	**−0.02 [−0.59, 0.56]**	**0.02 [−0.56, 0.59]**	**0.03 [−0.63, 0.68]**	**−0.09 [−0.45, 0.36]**

Effects are collapsed across all within-subjects variables: question type (Experiments 1, 2, and 3) and final assessment format (Experiment 4). 95% confidence intervals around *d* are in brackets.

that this should, in principle, lead to greater gains in long-term retention. In fact the results from Experiment 4 lend support for this theory, at least when the final test is in short-answer format. However our data show that the level of retrieval success is equally important, and that providing feedback is not always enough to make up for the lower levels of retrieval success on short-answer questions. When great effort to reduce the performance gap between the two formats is taken, then format differences emerge.

We did not see an advantage of combining short-answer and multiple-choice formats above using a simple short-answer test, even though those in the hybrid conditions answered each question twice (see also Butler et al., 2008). Perhaps subjects are simply remembering the answer during the second multiple-choice presentation rather than retrieving the answer (see Jacoby, 1978). If subjects had the opportunity to forget the answer to each individual question, making the second multiple-choice presentation within the hybrid format more difficult, then it is possible that retention would improve relative to the standard formats. Park's procedure is promising, but it is clear that future research will be needed to identify the best ways to use it to produce meaningful learning.

Even so, we believe using hybrid formats may be quite beneficial in educational settings. When using tests to implement retrieval practice, educators will need to balance requiring difficulty during retrieval while also making sure students are successful enough. Finding this sweet spot may be difficult, and it is likely going to be different for students with different levels of understanding of the material. Using hybrid formats allows the opportunity for more difficult or effortful retrieval, which should improve the learning outcomes of the retrieval practice, while still bringing students' success up using the multiple-choice questions. Our data suggest that this hybrid format will not be any worse than using other formats, and in some cases may produce more learning than a multiple-choice test would. As was mentioned in the introduction, multiple-choice questions are much easier for educators to score. The hybrid format retains this benefit—educators can score only the multiple-choice questions if they choose—while potentially improving learning outcomes.

We also did not find negative suggestion effects in three of our four experiments; multiple-choice questions did not cause students to produce the incorrect lures on the final assessment in Experiments 1, 2, and 4. In Experiment 3 we did find such a negative suggestion effect for those who were exposed to the incorrect alternatives on the multiple-choice questions. Importantly, this was the only experiment that did not provide direct feedback (i.e., the correct answer to the specific questions that were asked). Instead in Experiment 3 we provided students with the opportunity to restudy. Our pattern of results across four experiments suggests that not all forms of feedback are sufficient to avoid the negative consequences of multiple-choice testing. Instead, direct feedback in the form of the correct answer is likely necessary.

The present results show that both multiple-choice and short-answer question formats produce robust positive effects on long-term, meaningful learning, assessed with verbatim and inference questions. Contrary to their negative reputation, our results indicate that retrieval practice with multiple-choice questions can greatly improve meaningful learning. We consistently observed sizable effects of multiple-choice questions relative to the no retrieval practice control. Retrieval practice via short-answer questions with feedback has been recommended in order to maximise learning from tests (Kang et al., 2007). The use of feedback has been said to balance initial retrieval success differences between short-answer and multiple-choice formats. Our results indicate that retrieval success is an important factor, even when feedback is provided. Simply providing feedback will not always make up for lower levels of initial retrieval success. Importantly, feedback is not always provided in educational settings and, in the absence of feedback, multiple-choice questions are likely better for improving learning (see Kang et al., 2007). Computerised retrieval practice can be used in the classroom to help students learn and retain material for their courses, and hybrid formats can be used to balance retrieval difficulty and retrieval success.

REFERENCES

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 9, pp. 90–132). New York, NY: Academic Press.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency

is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.

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.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals* (Vol. 1). New York, NY: Longman.

Butler, A. C., Flanagan, P., Roediger, H. L., & McDaniel, M. A. (2007). *The benefit of generative study activities depends on the nature of the criterial test*. Poster presented at the Annual Meeting of the Psychonomic Society, Long Beach, CA.

Butler, A. C., Huelser, B. J., Caruso, C. A., & Roediger, H. L. (2008). *Examining Park's (2005) computer modified multiple-choice testing procedure*. Poster presented at the annual meeting of the Association for Psychological Science, Chicago, IL.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*, 941–956.

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.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review of quantitative synthesis. *Psychonomic Bulletin*, *132*, 354–380.

Clariana, R. B. (2003). The effectiveness of constructed-response and multiple-choice study tasks in computer aided learning. *Journal of Educational Computing Research*, *28*, 395–406.

Clariana, R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research & Development*, *49*, 23–36.

Cummings, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge Taylor & Francis Group.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, *6*, 217–226.

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

Frase, L. T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, *59*, 244–249.

Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, *17*, 45–50.

Gardiner, J. M., Craik, F. I. M., Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*, 213–216.

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, *56*, 212–242.

Haynie, W. J. (1994). Effects of multiple-choice and short-answer tests on delayed retention learning. *Journal of Technology Education*, *6*, 32–44.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649–667.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effects of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1250–1257.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.

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

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

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

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, *15*, 1–11.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*, 194–199.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., . . . Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*, 51–57.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399–414.

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*, 494–513.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalising test-enhancing learning from

the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200–206.
Morris, 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.
Park, J. (2005). Learning in a new computerised testing system. *Journal of Educational Psychology*, *97*, 436–443.
Park, J, & Choi, B. C. (2008). Higher retention after a new take-home computerised test. *British Journal of Educational Technology*, *39*, 538–547.
Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447.
Roediger, H. L., & Karpicke, J. D. (2006). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.
Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognitive*, *31*, 1155–1159.
Rogers, B. (2001). *TOEFL CBT Success*. Princeton, NJ: Peterson's.
Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/a0033569
Williams, J. P. (1963). Comparison of several response modes in a review program. *Journal of Educational Psychology*, *54*, 253–260.
Williams, J. P. (1965). Effectiveness of constructed-response and multiple-choice programing modes as a function of test mode. *Journal of Educational Psychology*, *56*, 111–117.

APPENDIX

Sample texts, and sample questions with response alternatives (multiple-choice format only). The correct response is typed in bold. (V) denotes a verbatim question; (I) denotes an inference question. The feedback corresponding to each question is also listed below.

Venice (first two paragraphs)

Venice is the capital of the region of Veneto and the province of the same name located in north-east Italy. At the heart of the city lies the ''Centro Historico'' (Historic Center), a mass of buildings and winding canals that has been inhabited since the 5th century AD. Roughly 62,000 people live in this neighbourhood that stretches across numerous small islands in the marshy Venetian Lagoon.

The majority of inhabitants, some 208,000 people, live on the land around the lagoon called ''Terraferma'' making the total population of the city much larger. In addition to bordering the Adriatic Sea, the saltwater lagoon stretches along the shoreline between the mouths of the Po and the Piave rivers. The abundance of water on all sides makes boats the primary mode of transportation for much of the citizens of Venice.

(V) What sea does the lagoon surrounding the city of Venice border?

1. **The Adriatic Sea**
2. The Sargasso Sea
3. The Baltic Sea
4. The Aegean Sea
5. The Black Sea

Feedback statement: The lagoon surrounding the city of Venice borders the Adriatic Sea.

(I) What is the total population of Venice?

1. 160,000 people
2. 530,000 people
3. 420,000 people
4. 350,000 people
5. **270,000 people**

Feedback statement: The total population of Venice is 270,000 people.

The Sun

The Sun today is a yellow dwarf star. It is fuelled by thermonuclear reactions near its centre that convert hydrogen to helium. The Sun has existed in its present state for about 4 billion, 600 million years and is thousands of times larger than the Earth.

By studying other stars, astronomers can predict what the rest of the Sun's life will be like. About 5 billion years from now, the core of the Sun will shrink and become hotter. The surface temperature will fall. The higher temperature of the centre will increase the rate of thermonuclear reactions. The outer regions of the Sun will expand approximately 35 million miles, which is about the distance to Mercury. The Sun will then be a red giant star. Temperatures on the Earth will become too hot for life to exist.

Once the Sun has used up its thermonuclear energy as a red giant, it will begin to shrink. After

it shrinks to the size of the Earth, it will become a white dwarf star. The Sun may throw off huge amounts of gases in violent eruptions called nova explosions as it changes from a red giant to a white dwarf.

After billions of years as a white dwarf, the Sun will have used up all its fuel and will have lost its heat. Such a star is called a black dwarf. After the sun has become a black dwarf, the Earth will be dark and cold. If any atmosphere remains there it will have frozen onto the Earth's surface.

(V) What type of star is the Sun today?

1. **Yellow dwarf star**
2. Red giant star
3. White dwarf star
4. Black dwarf star
5. Dark blue star

Feedback statement: Today the Sun is a yellow dwarf star.

(V) About 5 billion years from now, what two things will happen to the Sun's core?

1. **It will shrink and become hotter**
2. It will grow and become hotter
3. It will shrink and become cooler
4. It will grow and become cooler
5. It won't grow but will become hotter

Feedback statement: About 5 billion years from now, the Sun will shrink and become hotter.