APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **22**: 861–876 (2008)
Published online 19 September 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1391

Examining the Testing Effect with Open- and Closed-Book Tests

POOJA K. AGARWAL*, JEFFREY D. KARPICKE, SEAN H. K. KANG, HENRY L. ROEDIGER III and KATHLEEN B. McDERMOTT

Washington University in St. Louis, St. Louis, USA

SUMMARY

Two experiments examined the testing effect with open-book tests, in which students view notes and textbooks while taking the test, and closed-book tests, in which students take the test without viewing notes or textbooks. Subjects studied prose passages and then restudied or took an open- or closed-book test. Taking either kind of test, with feedback, enhanced long-term retention relative to conditions in which subjects restudied material or took a test without feedback. Open-book testing led to better initial performance than closed-book testing, but this benefit did not persist and both types of testing produced equivalent retention on a delayed test. Subjects predicted they would recall more after repeated studying, even though testing enhanced long-term retention more than restudying. These experiments demonstrate that the testing effect occurs with both open- and closed-book tests, and that subjects fail to predict the effectiveness of testing relative to studying in enhancing later recall. Copyright © 2007 John Wiley & Sons, Ltd.

A growing body of research has shown that taking a test can do more than simply assess learning: Tests can also enhance learning and improve long-term retention, a phenomenon known as the testing effect (see Carpenter, Pashler, & Vul, 2006; Karpicke & Roediger, 2007b; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006b). When subjects study and then take a test over material, they recall more on a delayed criterial test than if they had just studied the material once or if they had studied it repeatedly (see Roediger & Karpicke, 2006a, for review). The testing effect indicates that retrieval processes used when taking a test have powerful effects on learning and long-term retention. The fact that students engage in cognitive processes that promote learning when taking a test is often overlooked in education. Research aimed at understanding these processes has important implications for educational practice.

The purpose of the present research was to examine the testing effect with two types of tests commonly found in education: closed-book tests, the traditional method of testing students, and open-book tests, a method gaining popularity in primary, secondary and higher education (e.g. Baillie & Toohey, 1997; Ben-Chaim & Zoller, 1997; Eilertsen & Valdermo, 2000). Closed-book tests represent the norm, especially in higher education. During a closed-book test, students take the test without the aid of their notes or textbooks, and consulting

*Correspondence to: Pooja K. Agarwal, Washington University in St. Louis, Department of Psychology, Box 1125, One Brookings Drive, St. Louis, MO 63130-4899, USA. E-mail: pooja.agarwal@wustl.edu

supplementary material is typically considered cheating. In contrast, during open-book tests, students are allowed to consult their notes and textbooks while taking the test. Open-book tests are gaining favour among educators for a variety of reasons. For example, some educators believe that closed-book tests encourage students to engage in rote memorization when studying, whereas open-book tests encourage students to use higher-level thinking skills like problem solving and reasoning (Feller, 1994; Jacobs & Chase, 1992). In addition, students report that they experience less stress and anxiety when preparing for open-book tests than they do when preparing for closed-book tests (Theophilides & Dionysiou, 1996). For these reasons, some educators argue that open-book tests promote and assess learning more effectively than traditional closed-book tests (Cnop & Grandsard, 1994; Eilertsen & Valdermo, 2000; Theophilides & Koutselini, 2000).

Prior research has shown that tests enhance learning, but the effects of open- vs. closed-book tests on long-term retention have not been systematically investigated. There may be reasons to think that open-book tests might promote better learning than closed-book tests. For example, if open-book tests do encourage higher-level thinking skills and if practising these skills promotes long-term retention, then open-book tests may confer greater benefit than closed-book tests. Students might also commit fewer errors on open-book tests than they would on closed-book tests because open-book tests allow students to access answers during the test. Prior research on the testing effect has shown that if students make errors of commission on an initial test and do not receive corrective feedback, they may retain those errors on later tests and run the risk of incorporating false information into their general knowledge (see Butler, Marsh, Goode, & Roediger, 2006; Roediger & Marsh, 2005). Thus, open-book tests may benefit student learning because they might promote higher-level thinking (more than closed-book tests) and because they provide answers during the test so students make few, if any, errors of commission.

Alternatively, there are reasons why closed-book tests might enhance learning more than open-book tests. One theory of the testing effect holds that tests requiring more challenging retrieval produce greater benefits for long-term retention (see Bjork, 1999; Karpicke & Roediger, 2007a; McDaniel, Roediger, et al., 2007; Roediger & Karpicke, 2006a). Support for this idea comes from research comparing the testing effect with recall tests (which involve production of material, like short answer tests) and recognition tests (which involve identification of material, e.g. multiple-choice tests). The results of several studies converge on the conclusion that recall tests promote better long-term retention than recognition tests, regardless of whether the final criterial test requires recall or recognition (see Butler & Roediger, 2007; Glover, 1989; Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007). Other research has manipulated the difficulty of retrieval required on a test by having the test occur after a brief delay and the general finding is that delaying a first test produces positive effects on later retention (Jacoby, 1978; Karpicke & Roediger, 2007a; Modigliani, 1976; Pashler, Zarow, & Tripplett, 2003; Whitten & Bjork, 1977). These positive effects of challenging test conditions on long-term retention represent examples of Bjork's (1994, 1999) concept of creating desirable difficulties for learners. Conditions that require more difficult and challenging processing may slow initial learning but ultimately enhance long-term retention relative to less challenging learning conditions that produce rapid initial learning but poor retention.

Likewise, delaying feedback until after a test has also been shown to have positive effects on later retention (for review, see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). During open-book tests, students are able to receive immediate feedback about their

DOI: 10.1002/acp

performance. Closed-book tests, on the other hand, do not provide feedback during the test and any feedback occurs after students have completed the test at the earliest. Research using a variety of paradigms has shown that delayed feedback often promotes better long-term retention than providing immediate feedback. For instance, this principle has been demonstrated repeatedly in motor learning research (e.g. Schmidt, Young, Swinnen, & Shapiro, 1989). Butler, Karpicke, and Roediger (in press) showed that the positive effect of delayed feedback generalizes to educationally relevant tasks, as well. They found that delaying the presentation of feedback until after subjects had completed a multiple-choice test produced better long-term retention than presenting immediate feedback after each question on the test. Thus, although students might perform better on initial open-book tests than on closed-book tests because students have access to immediate feedback, closed-book tests may enhance long-term retention more than open-book tests due to the benefits of delayed feedback. Still, prior research on adjunct questions, pioneered by Rothkopf (1966), examines questions that are inserted either before (i.e. prequestions) or after (e.g. postquestions) a short passage. The typical finding in the adjunct question literature is that prequestions and postquestions result in similar retention of information (Anderson & Biddle, 1975; Rickards, 1979). Considering that open-book test questions may serve a similar purpose as prequestions (e.g. subjects can look over the questions before and during reading), while closed-book test questions are similar to postquestions, it remains unclear which test format promotes the greatest long-term retention.

The two experiments reported here examined the testing effect with open- and closed-book tests. In both experiments, students studied prose passages about a variety of educationally relevant topics (e.g. history, science and literature) and took short answer tests that were either closed-book (students took the test without viewing the passage) or open-book (students viewed the passage while taking the test). Experiment 1 investigated six conditions. Two conditions involved closed-book tests. In one closed-book test condition, subjects studied a passage one time and then took a short answer test. In another closed-book test condition, subjects studied the passage, took the short answer test and then received feedback by viewing the passage again and grading their test answers. Two other conditions involved open-book tests. In one open-book test condition, subjects read a passage and then took the short answer test while viewing the passage a second time. In another open-book test condition, subjects took the test and viewed the passage without a prior reading of the passage. We refer to this as the 'simultaneous answering' condition. The simultaneous answering condition is similar to what students might do if they answered practice questions (e.g. as if questions were embedded in a textbook chapter) without reading the text in advance. We also compared the open- and closed-book test conditions against a control condition in which subjects studied the passage but did not take an initial test, as well as against a non-studied control condition in which subjects only took a final test on the passage to measure baseline knowledge. Our primary interests were the effects of open- and closed-book tests on initial test performance and on long-term retention assessed after a week delay.

We expected to observe testing effects in both experiments; based on prior research we predicted that taking a test would enhance long-term retention more than studying the passage once. We also expected that providing feedback after an initial test would produce a positive effect on long-term retention (e.g. the closed-book test with feedback condition would outperform the closed-book test condition on the final test). Our primary goal was to determine whether open- or closed-book tests would be differentially or equally effective in producing a testing effect on the delayed final test.

DOI: 10.1002/acp

EXPERIMENT 1

Method

Subjects

Thirty-six subjects (26 females, ages 18–24 years old) were recruited from the Washington University in St. Louis Department of Psychology human subject pool. Subjects received either credit towards completion of a research participation requirement or cash payment.

Materials

Six prose passages, each approximately 1000 words in length ($M = 998.67$), were selected from a textbook used in education (Cooper et al., 1996). Each passage covered a single topic ('Arctic Explorer', 'Falcon', 'Fossils', 'Twisters', 'Voyager' and 'Wolves') and the average Flesch Reading Ease score for the six passages was 76.13 (Flesch, 1948). Each corresponding test consisted of seven short answer questions based on facts and ideas contained in each passage. For example, the following excerpt is from the passage about the Voyager spacecraft:

> The Voyagers were not very big—each one was about the size of a small car—but they were the most advanced spacecraft ever designed. The scientific instruments they carried included special cameras with telescopic lenses. These cameras would take close-up pictures of the giant planets and the surfaces of their moons. Other instruments would measure ultraviolet and infrared light. This light, invisible to normal cameras, would tell scientists more about the temperatures of the planets and what they are made of.

The corresponding test question asked, 'Why did the Voyager have instruments that would measure ultraviolet and infrared light?' The test questions were identical on the initial and final tests. Questions appeared on the test in the order in which the facts occurred in the passage.

Procedure

Table 1 shows the conditions used in the experiment. The six passages were presented in the same order for all subjects, but the order in which the six conditions occurred was counterbalanced. A Latin square was used to counterbalance the conditions, creating six counterbalancing orders and six subjects were assigned to each of the six orders. The six conditions were manipulated within-subjects. Once subjects completed one condition

Table 1. Learning conditions in Experiment 1

Condition	Session 1	Session 2
Study-once	Study	Test
Closed-book test	Study + test	Test
Closed-book test with feedback	Study + test + self-grade test (with passage available)	Test
Open-book test	Study + test (with passage available)	Test
Simultaneous answering	Study and test (with passage available) simultaneously	Test
Non-studied control		Test

Note: Session 2 occurred 1 week after Session 1. Subjects took a final closed-book test over each passage in Session 2.

DOI: 10.1002/acp

(e.g. studying and taking the test in the closed-book test condition), they moved on to the next condition, according to their counterbalancing order.

Subjects were tested individually or in small groups. In Session 1, subjects were instructed that they would read several prose passages, which might or might not be followed by a test. Thus, subjects did not know whether to anticipate a closed-book test, an open-book test or another passage. During a study period, subjects were told to read the passage and hand it back to the experimenter when they were finished; thus, study periods were self-paced. During a test, subjects were asked to write a response for every question, to be as detailed as possible, and to hand the test to the experimenter when complete; thus, test periods were also self-paced. The experimenter observed compliance with all instructions provided.

In the study-once condition, subjects read the passage one time and were not tested on it. In the closed-book test condition, subjects read the passage one time and then took a short answer test without viewing the passage again. In the closed-book test with feedback condition, subjects read the passage, took the short answer test without viewing the passage and then were given the passage and told to check their answers. Specifically, they were instructed to circle answers that were correct and to write an X through incorrect answers, without changing or adding to their original answers. In the open-book test condition, subjects read the passage one time and then were able to view the passage while taking the short answer test. In the simultaneous answering condition, subjects read the passage and completed the short answer test at the same time, but did not read the passage before taking the test. Finally, one passage was not studied in Session 1 but was tested in Session 2 to assess prior knowledge of the material on the final test.

Session 2 occurred 1 week after Session 1. In Session 2, subjects took final short answer tests over all six passages, without restudying or reviewing the passages (i.e. the final tests were closed-book). At the end of the experiment, subjects were debriefed and thanked for their time.

Results

Scoring

Subjects' responses on each test question were scored on a 3-point scale. Three points were awarded for detailed and complete answers; two points were given for correct but less detailed answers; one point was given for reasonable guesses that could have been drawn from prior knowledge rather than from passage content; and zero points were given for incorrect or blank answers (see Appendix for sample responses). Each test included seven questions, so the maximum score obtained on each test was 21 points. Two raters scored 10% of the tests and the Pearson product moment correlation between their scores was $r = .98$. Given the high inter-rater reliability, one rater scored the remaining tests. We also analysed the data by giving one point for correct answers and zero points for incorrect answers, and the pattern of results was the same by both scoring methods. Thus, we report analyses based on the 3-point scoring method.

Initial test performance

Table 2 shows the mean proportion correct on the initial tests and shows that subjects performed better on the open-book tests than on the closed-book tests, which is understandable. In fact, performance on the open-book tests were not perfect primarily because of the 3-point scoring system requiring detailed answers, which were not always provided. Errors of commission were rare (e.g. approximately 4% of all responses across

DOI: 10.1002/acp

Table 2. Mean proportion correct in Experiment 1

	Proportion correct	
Condition	Immediate test	One week delayed test
Study-once		.46
Closed-book test	.72	.59
Closed-book test with feedback	.69	.68
Open-book test	.81	.65
Simultaneous answering	.82	.63
Non-studied control		.18

conditions were incorrect answers). A one-way ANOVA revealed a main effect of the four test conditions on initial recall performance, $F(3, 105) = 11.27$, $\eta_p^2 = .24$. Combining the two open-book conditions and combining the two closed-book conditions, subjects performed better on the test when they could consult the passage (i.e. open-book test conditions, $M = .82$) than when they could not consult the passage (i.e. closed-book test conditions, $M = .70$), $t(35) = 5.87$, $d = 1.12$, $p_{rep} = 1.00$ (p_{rep} is an estimate of the probability of replicating the direction of an effect; see Killeen, 2005).

Final test performance

Table 2 also shows the mean proportion correct on the week-delayed criterial test for each condition. Performance in the non-studied control condition was relatively low ($M = .18$), far below all other conditions. A one-way ANOVA on the remaining five conditions revealed a main effect of learning condition on final recall performance, $F(4, 140) = 14.84$, $\eta_p^2 = .30$. Table 2 shows that both open- and closed-book tests produced large testing effects: All four test conditions led to better long-term retention than the study-once condition, ts$(35) > 3.48$, ds$ > .87$, $p_{reps} > .99$, confirming the testing effect under these disparate conditions. Performance in the closed-book test with feedback condition ($M = .68$) was greater than performance in the closed-book test without feedback condition ($M = .59$), $t(35) = 2.58$, $d = .57$, $p_{rep} = .94$, showing a positive effect of feedback on long-term retention. There was also a slight advantage of the open-book test condition ($M = .65$) relative to the closed-book test without feedback condition, $t(35) = 1.88$, $d = .45$, $p_{rep} = .90$. The open-book test and the closed-book test with feedback conditions resulted in similar final performance. Thus, while open-book tests resulted in substantially greater initial performance than closed-book tests, this boost in initial performance did not carry forward to the final test. Still, the results demonstrate robust testing and feedback effects, regardless of an open- or closed-book test format.

Discussion

In Experiment 1, taking an initial test after studying produced better long-term retention than studying without testing. Testing effects were observed with closed-book tests, when subjects did not view the passage while testing, and with open-book tests, when subjects were allowed to view the passage during the test. An effect of feedback was also present: the closed-book test with feedback condition resulted in greater final performance than the closed-book test without feedback condition. Although open-book tests led to better initial performance than closed-book tests, the open-book tests resulted in similar performance to the closed-book tests after a delay.

DOI: 10.1002/acp

EXPERIMENT 2

Experiment 2 had two main purposes. First, we wanted to replicate the testing and feedback effects observed in Experiment 1 and examine the effects of repeated studying on long-term retention. One criticism sometimes levelled against testing effect research is that when a study-test condition outperforms a study-once control condition, the testing effect could be due to re-exposure to the material rather than testing *per se*. Repeated study conditions alleviate this concern by re-exposing subjects to the entire set of the material, thus equating the number of times subjects are exposed to the material in the study and test conditions. Still, the testing effect occurs even with this more stringent control (see Hogan & Kintsch, 1971; Roediger & Karpicke, 2006a,b; Wheeler, Ewers, & Buonanno, 2003). In Experiment 2, we investigated conditions where students read a passage once (study 1×), twice (study 2×) or three times (study 3×). The number of exposures in the study 2× condition was equal to the number of exposures in the open- and closed-book test conditions. Likewise, exposure to material in the study 3× condition was equal to the number of exposures in the closed-book test with feedback condition. For all test and relevant study control conditions, we predicted that testing would lead to better long-term retention than repeated studying.

The second purpose of Experiment 2 was to examine subjects' metamemorial abilities following repeated studying or testing in open- or closed-book test conditions. In Session 1, after the last period in each condition (i.e. following the last study period in the repeated study conditions, or after the test/feedback period in the test conditions), we asked subjects to predict how well they would remember the passage on a final test in 1 week (i.e. to provide an aggregate judgment of learning). We expected to find differences in judgments of learning (JOLs) in the study and test conditions, based on the idea that subjects rely on different cues when making JOLs after studying or testing (Dunlosky & Nelson, 1992; Koriat, 1997; Roediger & Karpicke, 2006b). In particular, prior research showed that after studying, subjects base their JOLs on the difficulty of the material or learning task, whereas following testing, subjects base their JOLs on the subjective likelihood of recalling an item. We also examined differences in JOLs in the open- and closed-book test conditions, as subjects may use different cues when making JOLs following these types of tests. JOLs bear practical importance because they indicate what study strategies students choose and guide students' allocation of subsequent study-time (see Kornell & Metcalfe, 2006; Nelson & Narens, 1994; Son & Metcalfe, 2000).

Experiment 2 examined eight conditions (see Table 3). Two involved restudying without testing (the study 2× and study 3× conditions). The remaining six conditions (a study 1× condition, four test conditions and a non-studied control condition) were the same as those used in Experiment 1. After the last period in each condition during Session 1, subjects made an aggregate prediction of how well they would remember the passage on a final test in 1 week. One week later, subjects took final short answer tests over all of the passages (as they did in Experiment 1).

Method

Subjects

Forty-eight subjects (32 females, ages 18–23 years old) were recruited from the Washington University in St. Louis Department of Psychology human subject pool. These

DOI: 10.1002/acp

Table 3. Learning conditions in Experiment 2

Condition	Session 1	Session 2
Study 1×	Study	Test
Study 2×	Study + study	Test
Study 3×	Study + study + study	Test
Closed-book test	Study + test	Test
Closed-book test with feedback	Study + test + self-grade test (with passage available)	Test
Open-book test	Study + test (with passage available)	Test
Simultaneous answering	Study and test (with passage available) simultaneously	Test
Non-studied control		Test

Note: At Session 2, 1 week after Session 1, all subjects received a final closed-book criterial test.

subjects did not participate in Experiment 1 and they received either credit towards completion of a research participation requirement or cash payment.

Materials

Eight passages were used in Experiment 2. The six passages used in Experiment 1 were used again, and two additional passages (titled 'Earthquakes' and 'Santa Fe Trail') were selected from the same source as the other passages. The average Flesch Reading Ease score for the eight passages was 74.63 (Flesch, 1948). Each passage was approximately 1000 words in length ($M = 1001.25$), and short answer tests were constructed by the same method used in Experiment 1.

Procedure

The procedure was similar to that used in Experiment 1. Table 3 shows the within-subjects conditions used in Experiment 2. The eight passages were presented in the same order for all subjects, but the order in which the eight conditions occurred was counterbalanced. A Latin square was used to determine eight counterbalance orders and six subjects were assigned to each of the eight orders. Once subjects completed one condition (e.g. studying and taking the test in the closed-book test condition), they moved on to the next condition, according to their counterbalancing order. Subjects did not know what to expect after reading a passage for the first time (i.e. whether they would re-read the passage, take a closed-book test, or take an open-book test). The experimenter observed compliance with all instructions provided.

Subjects were tested individually or in small groups. Three of the passages were studied but not tested in Session 1. In the study 1× condition, subjects read the passage one time. In the study 2× condition, subjects read the passage two times and in the study 3× condition, they read the passage three times. The four test conditions (two open- and two closed-book test conditions) and the non-studied control condition were identical to those used in Experiment 1.

After the last study/test period in each condition, subjects made an aggregate JOL by predicting how well they would remember the passage after a delay. Specifically, subjects were asked, 'How well do you think you will remember this passage in 1 week?' Subjects made their judgments using a 0–100% scale, where 0% meant that they did not think they would remember anything from the passage in 1 week, 100% meant they thought they

DOI: 10.1002/acp

would remember the passage perfectly in 1 week and intermediate values reflected intermediate levels of confidence.

Session 2 occurred 1 week after Session 1. In Session 2, subjects took final short answer tests over all eight passages without restudying or reviewing the passages (i.e. the final tests were closed-book). Final short answer test questions were the same as initial short answer test questions. At the end of the experiment, subjects were debriefed and thanked for their time.

Results

The 3-point scoring method used in Experiment 1 was used again in Experiment 2. Two raters scored 10% of the tests and the correlation between their scores was $r = .97$. Given the high inter-rater reliability, one rater scored the remaining tests.

Initial test performance

Table 4 shows the mean proportion correct on the immediate tests and shows that subjects performed better on the initial open-book tests than on the initial closed-book tests, as in Experiment 1. A one-way ANOVA confirmed that there was a main effect of learning condition on immediate recall performance, $F(3, 141) = 28.70$, $\eta_p^2 = .38$. Combining the two open-book conditions and combining the two closed-book conditions, subjects performed better on the initial test when they viewed the passage in the open-book test conditions ($M = .82$) than when they could not view the passage in the closed-book test conditions ($M = .66$), $t(47) = 7.56$, $d = 1.41$, $p_{rep} = 1.00$.

Final test performance

Table 4 also shows the mean proportion correct on the week-delayed criterial tests. Performance in the non-studied control condition was, again, relatively low ($M = .16$) and all other conditions showed learning relative to this baseline condition. A one-way ANOVA on the remaining seven conditions demonstrated a main effect of learning condition on final recall performance, $F(6, 282) = 27.49$, $\eta_p^2 = .37$. The top portion of Table 4 shows a positive effect of repeated studying on long-term retention: Performance in the study 2× ($M = .50$) and the study 3× ($M = .54$) conditions was greater than performance in the study 1× condition ($M = .40$), $ts(47) > 3.95$, $ds > .65$, $p_{reps} = 1.00$.

Testing enhanced long-term retention more than restudying, and the test conditions outperformed their relevant study control conditions. Similar to Experiment 1, all four testing conditions led to greater final performance than the study 1× condition,

Table 4. Mean proportion correct and mean judgments of learning (JOLs) in Experiment 2

	Proportion correct		
Condition	Immediate test	One week delayed test	JOL
Study 1×		.40	.57
Study 2×		.50	.65
Study 3×		.54	.71
Closed-book test	.67	.55	.62
Closed-book test with feedback	.65	.66	.66
Open-book test	.81	.66	.65
Simultaneous answering	.83	.59	.65
Non-studied control		.16	

DOI: 10.1002/acp

$ts(47) > 5.64$, $ds > .94$, $p_{reps} = 1.00$. The open-book ($M = .66$) and closed-book ($M = .55$) test conditions (in which subjects were exposed to the material twice) led to better recall than the study 2× condition ($M = .50$), $ts(47) > 2.12$, $ds > .37$, $p_{reps} > .93$. Similarly, performance in the closed-book test with feedback condition (in which subjects were exposed to the material three times, $M = .66$) was greater than performance in the study 3× condition ($M = .54$), $t(47) = 4.95$, $d = .81$, $p_{rep} = 1.00$. Furthermore, simply completing an open-book test without ever studying in the first place (as in the simultaneous answering condition, $M = .59$) led to better performance than studying once ($M = .40$), $t(47) = 8.39$, $d = 1.21$, $p_{rep} = 1.00$ or even three times ($M = .54$), $t(47) = 2.04$, $d = .39$, $p_{rep} = .92$.

In addition, the closed-book test with feedback and the open-book test conditions produced better performance than the closed-book test condition, $ts(47) > 3.91$, $ds > .70$, $p_{reps} = 1.00$, showing a positive effect of feedback on retention. Again, although subjects performed better on an initial open-book test, final performance in the open-book test condition was similar to final performance in the closed-book test with feedback condition. Both conditions, however, resulted in greater performance than the simultaneous answering condition, $ts(47) > 2.70$, $ds > .43$, $p_{reps} > .97$.

Judgments of Learning

Finally, Table 4 shows subjects' mean JOLs made after the last period in Session 1 in each condition. Overall, subjects' JOLs increased with repeated studying. Subjects predicted that they would recall more after studying twice ($M = .65$) than after studying once ($M = .57$), $t(47) = 2.77$, $d = .41$, $p_{rep} = .97$, and that they would recall more after studying three times ($M = .71$) than after studying twice, $t(47) = 3.13$, $d = .36$, $p_{rep} = .98$. In contrast, JOLs did not differ across the four test conditions. A one-way ANOVA confirmed that there were no differences among the open- and closed-book test conditions ($F < 1$).

A closer examination of predicted (JOLs) and actual final recall performance shows a striking dissociation: When equating the number of presentations, JOLs were greater after repeated studying than after testing following two or three exposures to material, even though final recall was greater after testing than after restudying in the relevant comparison conditions. Figure 1 shows JOLs and final recall in conditions in which subjects were

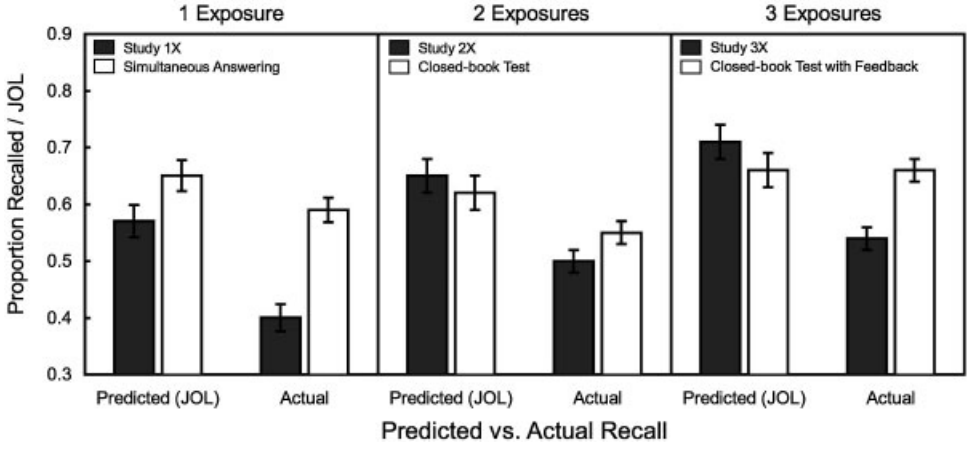

Figure 1. Predicted vs. actual final test performance as a function of the number of exposures in Experiment 2. Error bars represent standard errors of the mean

DOI: 10.1002/acp

exposed to the passages once (the study 1× and simultaneous answering conditions), two times (the study 2× and closed-book test conditions) and three times (the study 3× and closed-book test with feedback conditions). The figure shows that after one exposure to the material, subjects predicted they would recall more in the simultaneous answering condition than in the study 1× condition, consistent with actual final test performance. Following more than one exposure to material, however, subjects predicted they would recall more in the restudy conditions than in the test conditions, inconsistent with actual final test performance since testing enhanced long-term retention more than repeated studying. The results in Figure 1 indicate that subjects have little metacognitive awareness of the testing effect, a point to which we return in the General Discussion.

Discussion

Experiment 2 showed that both an open-book test and a closed-book test with feedback enhanced long-term retention more than studying once, repeated studying or testing without feedback, replicating testing and feedback effects observed in Experiment 1 and in previous research. In addition, open-book tests produced better initial performance than closed-book tests, but open-book tests led to similar performance relative to closed-book tests after a delay. Finally, perhaps most striking were the differences in JOLs following study and test conditions: Subjects predicted they would recall more after repeated studying than after testing, even though testing enhanced long-term retention more than restudying.

GENERAL DISCUSSION

The primary results of our research can be summarized as follows. First, both open- and closed-book tests produced a testing effect: Taking a test enhanced long-term retention more than studying the passage once, and Experiment 2 showed that testing produced better long-term retention than repeated studying, replicating previous findings (Roediger & Karpicke, 2006b). Second, powerful effects of feedback were demonstrated in both experiments, such that tests with feedback (e.g. the open-book test and the closed-book test with feedback conditions) outperformed the closed-book test without feedback condition. Third, an initial benefit from open-book tests did not last over a 1 week delay; instead, the open-book test and the closed-book test with feedback conditions resulted in similar final performance in both experiments. Finally, in Experiment 2, when subjects were asked to predict how well they would remember material in the future, they predicted that they would remember the passages better after repeatedly studying them than after testing on them, even though the opposite result was true. We focus our discussion on the latter two results: (1) An initial benefit from open-book tests does not last after a delay and (2) subjects predict that they will recall more after restudying than after testing, in opposition to their actual recall performance.

Although we consistently found that performance was better on initial open-book tests than on closed-book tests, this outcome may not occur in real world educational settings in which students regulate how they study based on the type of test they expect. In our experiments, subjects did not know whether to expect an upcoming open- or closed-book test after studying the passage. Prior research suggests that students who expect an open-book test may study less (or less effectively) than those who expect a closed-book test

DOI: 10.1002/acp

(Ioannidou, 1997). Other classroom studies have also reported that students sometimes perform worse on open-book tests than on closed-book tests because they prepare less effectively when they will be allowed to use their notes or textbooks during the test (Boniface, 1985; Kalish, 1958; Pauker, 1974; Weber, McBee, & Krebs, 1983). Thus, in comparison to closed-book tests, open-book tests may have a negative effect of reducing the effectiveness of students' studying.

In addition, we suspect that any positive effects of closed-book tests may be even more powerful in a repeated testing design. For example, the challenging processing required on two initial closed-book tests with feedback (i.e. study—closed-book test—feedback—closed-book test—feedback) may produce much better long-term retention than having two initial open-book tests (e.g. study—open-book test—open-book test), even though performance would be at ceiling on the second test in both conditions. Of course, these speculations await further research.

Our results in the current study lend support to the idea that challenging retrieval processes promote long-term retention (see too Bjork, 1999; Chan, McDermott, & Roediger, 2006; McDaniel, Roediger, et al., 2007; Roediger & Karpicke, 2006a). Both open-book and closed-book tests can be considered a desirable difficulty (Bjork, 1994, 1999) in that they require more difficult, challenging processing than restudying a passage, yet this difficult processing benefits long-term retention. The results can also be explained by a theory of retrieval difficulty proposed by Bjork and Bjork (1992). They argued that information in memory might be characterized in terms of two strengths: Storage strength, which refers to a permanent property of the information that determines long-term retention, and retrieval strength, which refers to the momentary accessibility of the information. In their theory, increments in retrieval strength are negatively correlated with increments in storage strength. That is, when retrieval strength is high and information is easily accessible, the retrieval of that information produces small increments in storage strength. In contrast, lower retrieval strength and more difficult retrieval produce greater increments in storage strength and thereby promote long-term retention, assuming the item can be retrieved. In the present experiments, open-book test conditions increased the momentary accessibility of some of the information, as evidenced by high initial performance in these conditions. The retrieval of accessible information on the initial open-book tests, however, produced small increments in storage strength. This resulted in similar long-term retention following open-book tests relative to retention following closed-book tests.

Turning now to our metamemory results, Experiment 2 showed that subjects predicted they would recall more after repeated studying than after testing (i.e. after two or three exposures to material), even though the opposite was true and testing enhanced long-term retention. Prior research has revealed similar results. Roediger and Karpicke (2006b) showed that students predicted they would recall more on a delayed final test after studying a passage in four study periods than after studying once and taking three free recall tests. Karpicke, McCabe, and Roediger (2007) extended this finding by showing that subjects are generally overconfident and fail to predict forgetting over a week delay after repeated studying, whereas they are underconfident and do take forgetting into account after repeated testing (cf. Koriat et al., 2004). Thus, the emerging result in the testing effect literature is that although testing enhances long-term retention, students generally lack metacognitive awareness of the testing effect (see too Karpicke, Butler, & Roediger, 2007).

The different patterns of JOLs following studying and testing can be explained within the cue-utilization framework proposed by Koriat (1997). Koriat offered that JOLs are inferential in nature and can be based on a variety of cues available in a given context. He

DOI: 10.1002/acp

distinguished among three broad classes of cues: Intrinsic cues, which are properties of the materials that disclose their inherent ease or difficulty; extrinsic cues, which are properties of the learning task such as the number of times an item was studied; and internal mnemonic cues, which are subjective indices of the likelihood an item will be recalled such as encoding or retrieval fluency. Koriat argued that a test during learning leads subjects to shift from reliance on intrinsic or extrinsic cues to utilization of mnemonic cues (see too Koriat, Sheffer, & Ma'ayan, 2002). In the present study, the results of Experiment 2 are consistent with the cue-utilization framework in that subjects relied on intrinsic or extrinsic cues when making JOLs after studying, but they used internal mnemonic cues to make JOLs after testing. Of course, differences in JOLs have important practical implications because students often base decisions about what to study and when to stop studying on their subjective assessment of their own learning. Clearly, JOLs differ greatly when made at study or at test.

In sum, although open-book tests are gaining popularity and result in large initial benefits, they produce a similar amount of long-term retention as traditional closed-book tests. Still, both types of tests enhance learning more than restudying or testing without feedback, thus open-book and closed-book tests should be used as strategies to improve the retention of material. In addition, subjects predicted they would recall more in the future after repeated studying than after testing, even though the opposite was true and testing enhanced long-term retention relative to restudying. This metacognitive illusion could have dire consequences for students when they must monitor and regulate their own learning.

ACKNOWLEDGEMENTS

This research was supported by Hoopes Undergraduate Research Awards from the Washington University Undergraduate Research Office, and grants from the Institute of Education Sciences and the James S. McDonnell Foundation. We thank Jessica Logan for her assistance in scoring recall tests and Jane McConnell for her help. We also thank Mark McDaniel, Jason Chan, Karl Szpunar and Andrew Butler for their insightful comments and valuable discussions. Experiment 1 was conducted as part of an undergraduate honours thesis completed by Pooja K. Agarwal, portions of which were presented at the 18th Annual Meeting of the Association for Psychological Science, New York City, New York, May 2006, and at the 47th Annual Meeting of the Psychonomic Society, Houston, Texas, November 2006.

REFERENCES

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *Psychology of learning and motivation* (Vol. 9, pp. 89–132). New York, NY: Academic Press.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213–238.

Baillie, C., & Toohey, S. (1997). The ''power test:'' Its impact on student learning in a materials science course for engineering. *Assessment & Evaluation in Higher Education*, *22*, 33–49.

Ben-Chaim, D., & Zoller, U. (1997). Examination-type preferences of secondary school students and their teachers in the science disciplines. *Instructional Science*, *25*, 347–367.

DOI: 10.1002/acp

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.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Boniface, D. (1985). Candidates' use of notes and textbooks during an open-book examination. *Educational Research*, *27*, 201–209.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (in press). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied.*

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*, 10.1080/09541440701326097.

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.

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.

Cnop, I., & Grandsard, F. (1994). An open-book exam for non-mathematics majors. *International Journal of Mathematical Education in Science and Technology*, *25*, 125–130.

Cooper, J. D., Pikulski, J. J., Au, K. H., Calderón, M., Comas, J. C., Lipson, M. Y., et al. (1996). *Explore*. Boston, MA: Houghton Mifflin Company.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374–380.

Eilertsen, T. V., & Valdermo, O. (2000). Open-book assessment: A contribution to improved learning? *Studies in Educational Evaluation*, *26*, 91–103.

Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation*, *20*, 235–238.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

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

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, *10*, 562–567.

Ioannidou, M. K. (1997). Testing and life-long learning: Open-book and closed-book examination in a university course. *Studies in Educational Evaluation*, *23*, 131–139.

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

Jacobs, L. C., & Chase, C. I. (1992). *Developing and using tests effectively: A guide for faculty.* San Francisco, CA: Jossey-Bass Publishers.

Kalish, R. A. (1958). An experimental evaluation of the open book examination. *Journal of Educational Psychology*, *49*, 200–204.

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

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2007). Test-enhanced learning, metacognition, and technology. Manuscript in preparation.

Karpicke, J. D., McCabe, D. P., & Roediger, H. L. (2007). Testing enhances recollection: Process dissociations and metamemory judgments. Manuscript in preparation.

Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval 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.

DOI: 10.1002/acp

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345–353.
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.
Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*, 643–656.
Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*, 147–162.
Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 609–622.
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). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200–206.
Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning & Memory*, *2*, 609–622.
Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.
Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057.
Pauker, J. D. (1974). Effect of open book examinations on test performance in an undergraduate child psychology course. *Teaching of Psychology*, *1*, 71–73.
Rickards, J. P. (1979). Adjunct postquestions in text: A critical review of methods and processes. *Review of Educational Research*, *49*, 181–196.
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.
Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159.
Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, *3*, 241–249.
Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 352–359.
Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.
Theophilides, C., & Dionysiou, O. (1996). The major functions of the open-book examination at the university level: A factor analytic study. *Studies in Educational Evaluation*, *22*, 157–170.
Theophilides, C., & Koutselini, M. (2000). Study behavior in the closed-book and open-book examination: A comparative analysis. *Educational Research and Evaluation*, *6*, 379–393.
Weber, L. J., McBee, J. K., & Krebs, J. E. (1983). Take home tests: An experimental study. *Research in Higher Education*, *18*, 473–483.
Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.
Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior*, *16*, 465–478.

DOI: 10.1002/acp

APPENDIX

Sample test responses

The following is an excerpt from the passage about the Voyager spacecraft:

> The Voyagers were not very big—each one was about the size of a small car—but they were the most advanced spacecraft ever designed. The scientific instruments they carried included special cameras with telescopic lenses. These cameras would take close-up pictures of the giant planets and the surfaces of their moons. Other instruments would measure ultraviolet and infrared light. This light, invisible to normal cameras, would tell scientists more about the temperatures of the planets and what they are made of.

The corresponding test question asked, 'Why did the Voyager have instruments that would measure ultraviolet and infrared light?'

Example of a response to the above question receiving three points for a detailed and complete answer: 'The Voyager had instruments that would measure ultraviolet and infrared light in order to tell scientists about planet temperatures and composition'.

Example of a response receiving two points for a correct but less detailed answer: 'The Voyager had instruments that would measure light in order to tell scientists about planet temperatures'.

Example of a response receiving one point for a reasonable answer that could have been drawn from prior knowledge: 'The Voyager had instruments that would measure light in order to provide more information about the planet'.

DOI: 10.1002/acp