APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **23**: 760–771 (2009)
Published online 26 August 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1507

Using Tests to Enhance 8th Grade Students' Retention of U.S. History Facts[†,‡]

SHANA K. CARPENTER[1,2]*,
HAROLD PASHLER[2] and NICHOLAS J. CEPEDA[2,3]

[1]*Iowa State University, USA*
[2]*University of California, USA*
[3]*York University, Canada*

SUMMARY

Laboratory studies show that retention of information can be powerfully enhanced through testing, but evidence for their utility to promote long-term retention of course information is limited. We assessed 8th grade students' retention of U.S. history facts. Facts were reviewed after 1 week, 16 weeks or not reviewed at all. Some facts were reviewed by testing (Who assassinated president Abraham Lincoln?) followed by feedback (John Wilkes Booth), while others were re-studied. Nine months later, all students received a test covering all of the facts. Facts reviewed through testing were retained significantly better than facts reviewed through re-studying, and nearly twice as well as those given no review. The best retention occurred for facts that were reviewed by testing after a 16-week time interval. Although the gain in item was numerically small, due to floor effects, these results support the notion that testing can enhance long-term retention of course knowledge. Copyright © 2008 John Wiley & Sons, Ltd.

In educational contexts, tests are commonly used as assessment devices to measure how much students have learned. Findings from experimental studies of memory suggest that tests are useful for more than just assessment, however. Many studies report that taking a test over some material actually renders that material more likely to be successfully remembered in the future (e.g. Allen, Mahler, & Estes, 1969; Izawa, 1992; McDaniel & Masson, 1985; Wheeler & Roediger, 1992). Taking a test is more effective in promoting future retention even when compared to spending additional time re-reading the material (e.g. Carpenter & DeLosh, 2005; Carpenter, Pashler, & Vul, 2006; Carrier & Pashler, 1992; Kuo & Hirshman, 1996, 1997). Furthermore, recent data have shown that testing, compared to restudying, actually reduces the rate at which information is forgotten from memory over time (Carpenter, Pashler, Wixted, & Vul, 2008; see also Roediger &

*Correspondence to: Shana K. Carpenter, Department of Psychology, Iowa State University, W112 Lagomarcino Hall, Ames, IA 50011-3180, USA. E-mail: scarpenter@ucsd.edu
[†]The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.
[‡]Parts of this study were presented at the 2007 All Hands Meeting of the Temporal Dynamics of Learning Center, La Jolla, 7–9 January, the 2007 Annual Meeting of the Cognitive Science Society, Nashville, 1–4 August and at the 2007 Annual Meeting of the Psychonomic Society, Long Beach, 15–18 November.

Karpicke, 2006a). The use of tests as learning devices would seem to be one concrete tool that educators can use to improve students' learning and reduce their forgetting.

There has been increasing interest in applying the findings from experimental studies of memory to enhancing acquisition and retention of knowledge in educational domains (e.g. Dempster, 1996; Kornell & R. A. Bjork, 2007; McDaniel, 2007; McDaniel, Roediger, & McDermott, 2007b; Metcalfe & Kornell, 2007; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Viadero, 2006). However, the benefits of testing for direct enhancement of learning do not appear to be widely recognized in educational practice. One reason for this could be the paucity of studies aimed at exploring the benefits of testing in real educational environments. The majority of studies on the testing effect have been conducted in the laboratory using relatively simple materials such as word lists (e.g. for a review, see Roediger & Karpicke, 2006b).

Recently, several noteworthy studies have demonstrated robust benefits of testing using educationally relevant materials such as text passages (e.g. Agarwal, Karpicke, Kang, Roediger, & McDermott, in press; Chan, McDermott, & Roediger, 2006; Marsh, Roediger, R. A. Bjork, & E. L. Bjork, 2007; Roediger & Marsh, 2005), scientific articles (e.g. Kang, McDermott, & Roediger, 2007), general knowledge questions (e.g. McDaniel & Fisher, 1991), English vocabulary learning (e.g. Cull, 2000), foreign language vocabulary learning (e.g. Carrier & Pashler, 1992; Pashler et al., 2007) and learning of visuospatial information in maps (Carpenter & Pashler, 2007). These studies offer some encouragement about the potential of tests to enhance learning across a wide variety of materials of the type that students often encounter in school.

As of yet, however, there have been few efforts to explore the effects of testing for promoting students' retention of their course material in actual classroom environments. Among the few studies that have been conducted in classroom environments, nearly all involved college-level courses (e.g. Bangert-Drowns, Kulik, & Kulik, 1991; for a recent review of both classroom and laboratory studies of the testing effect, see Roediger & Karpicke, 2006b). For example, in one study, McDaniel, Anderson, Derbish, and Morrisette (2007a) found that final exam scores in an online Brain and Behaviour course were higher when students reviewed the information through quizzes as opposed to additional re-reading.

Nearly everything we know about the testing effect has been derived from the way that adult college students learn information in relatively well-controlled environments—most often, in the laboratory. We would be on stronger ground in advocating widespread use of testing at all levels of education if we had evidence showing their efficacy within more diverse educational settings. Students in middle school or high school, for example would seem more likely than college students to represent a wide range of social and economic backgrounds, level of interest in the material, motivation and study habits. Direct assessment of the influence of these background factors was not the purpose of the current study, nor has it been addressed in any of the past work on testing effects that we are aware of. The purpose of the current investigation was to explore the robustness of the testing effect beyond the typical experimental setting on which the majority of research has so far been based.

The current study extended beyond the known research in this area by exploring testing benefits with actual course material over much longer retention intervals than have been explored in past studies. In one study, for example, Sones and Stroud (1940) found that retention of 7th grade students after 42 days was better when information was reviewed after 1 and 3 days in the form of a multiple choice test, as opposed to re-reading. Spitzer (1939) found that retention of 6th grade students after 63 days was enhanced when information was reviewed through testing as opposed to not being reviewed at all (Spitzer's

DOI: 10.1002/acp

study did not directly compare the effects of testing to re-reading, however). Furthermore, the materials used in these previous studies consisted of novel materials that did not pertain to information that students were currently learning in their courses. The current study, on the other hand, explored potential benefits of testing on 8th grade students' retention of their course material over a substantial time interval of 9 months—more than 250 days.

We gave students a review over facts they learned in their U.S. history course. Some facts were reviewed through testing with feedback (Test/Study), other facts were reviewed through simply re-reading (Study), and other facts were not reviewed at all (No Review). In classroom settings, instructors would seem to have some degree of control over when to schedule a review. In order to promote long-term retention, should the review occur relatively soon after learning, or after some time has passed?

Some laboratory studies have reported that final retention benefits more from a test that is given after a delay of several seconds or minutes, compared to a test given immediately after learning the material (e.g. Izawa, 1992; Whitten & R. A. Bjork, 1977). In the current study, we explored these effects over much longer and more educationally realistic time intervals. Half of the students (Immediate Review Group) received the review 1 week after learning the material, and the other half (Delayed Review Group) received the review after 16 weeks. After 36 weeks (i.e. 9 months), all students completed a final test over the facts. Students were tested over the same facts that they previously reviewed through Test/Study vs. Study, as well as the facts that were given No Review.

METHOD

Subjects

Students were recruited from a charter school in San Diego, CA, that enrols students in grades 6 through 12. Students were drawn from a population of 8th graders enrolled in one of five sections of U.S. history, each of which contained approximately 15–20 students. Each of these class sections was taught by one of two instructors. From these classes, a total of 75 students (44 females and 31 males) agreed to participate.

Each class section was randomly assigned to either the Immediate Review Group ($n = 37$), or the Delayed Review Group ($n = 38$). Due to constraints imposed by the practical requirements of administering the study in this particular school, it was not feasible to assign students to groups on an individual-student basis, as would have been optimal. However, we were able to ensure that each instructor had at least one class section assigned to the Immediate Review Group, and one to the Delayed Review Group. Any effects of instructor, therefore, were counterbalanced across groups.

During the course of the study (lasting approximately 1 year), five students from the Immediate Review Group ceased attending the school, three from the Delayed Review Group were absent on the day of the review and five (three from the Immediate Review Group, and two from the Delayed Review Group) were absent on the day of the final test. All analyses were based on the 29 remaining students in the Immediate Review Group, and the 33 remaining students in the Delayed Review Group.

Materials

The course material used in the study was collected by an undergraduate research assistant who attended each class for approximately 3 weeks while students were studying a unit on

DOI: 10.1002/acp

slavery and sectionalism. Questions for the study were constructed from material that had been covered in class discussions, notes, reading assignments and handouts pertaining to this unit. A set of 45 questions was created, each requiring a brief answer over some factual content (e.g. *Who assassinated president Abraham Lincoln?*). A complete list of the stimuli can be obtained from the authors upon request. For each student, 15 of these questions were randomly assigned to be reviewed through testing with feedback (Test/Study), 15 through re-studying (Study) and 15 were not reviewed at all (No Review).

Design and procedure

Each student was given a consent form packet containing a letter addressed to parents/guardians explaining the study, a form to be signed by the parent or guardian indicating permission for the student to participate, and a similar form to be signed by the student. Students were told that if they chose to participate in the experiment, they would be tested over information pertaining to 8th grade U.S. history. Seventy-five students and their parents agreed to have them participate and six declined.

All students received one review session and one final test session that entailed providing brief written answers to the questions that were created for the study. Students in the Immediate Review Group received the review session about 1 week after they completed their exams and standardized assessments in the course, whereas students in the Delayed Review Group received the review after 16 weeks. Thus, all students finished their required coursework prior to beginning the study. Students were not required to take U.S. history again until the 11th grade, and so it seemed unlikely that, during the course of the study, they would encounter much additional instruction over the information.

Review session

A researcher visited each class during the scheduled time for the review. Each student was given a sheet of paper containing 30 questions (e.g. *Who assassinated president Abraham Lincoln?*). For 15 of these, the answer was provided (e.g. *John Wilkes Booth*), whereas for the other 15, the answer was not provided. If the answer was provided (i.e. a Study item), students were asked to read both the question and answer. If an answer was not provided (i.e. a Test/Study item), students were asked to read the question and write down an answer, and were permitted to leave it blank if they were uncertain. Each student received a unique review sheet with a different set of 15 items randomly assigned to Study, 15 items randomly assigned to Test/Study, and 15 items randomly assigned to the No Review condition. The 30 items that appeared on the review (15 Study and 15 Test/Study) occurred in a different random order for each student.

Students were told that their participation was confidential so they should not write their name anywhere on the review materials, and later examination of the materials indicated that all students complied with this instruction. Students were asked to complete the review individually and then return it to the researcher, who would then give them an answer sheet. Each student's answer sheet was individually created to contain the same 30 items that appeared on their review, but the 30 items now appeared in alphabetical order. The new ordering was intended to increase the likelihood that students would read through each item again. Students were instructed to read the questions and answers for all 30 items, and to mark any of the 15 Test/Study items that they answered incorrectly. When finished, students returned their materials to the researcher. No time limit was imposed for the review, and most students completed it within about 20 minutes.

DOI: 10.1002/acp

Final test

All students received the final test 36 weeks after they completed the review. To ensure that the same amount of time elapsed in between the review and final test for both groups, the Delayed Review Group (who received their review 16 weeks after the Immediate Review Group) received their final test 16 weeks after the Immediate Review Group. The final test contained the same 15 Study items and 15 Test/Study items that appeared on the review, along with 15 items pertaining to 8th grade U.S. history that never appeared on the review (i.e. No Review items). Thus, the format of the test questions was the same on the review and final test, as is typical in many studies of the testing effect (e.g. Carpenter & DeLosh, 2006; Roediger & Karpicke, 2006b). Designing the study in this way also allowed the same scoring system to be applied for students' answers on the review and final test (see the Subsection 'Scoring').

A researcher visited each class to administer the final test. Students were asked to write down an answer to each of the 45 questions and were encouraged to guess if they were uncertain. The researcher reminded students that their participation was confidential so they should not write their name anywhere on the final test, and later examination of the materials indicated that all students complied with this instruction. Students were asked to complete the test individually and then return it to the researcher when finished. No time limit was imposed, and most students completed the final test within about 15 minutes. Upon finishing the final test, students were thanked for their participation. The researcher then provided the teachers with written debriefing information to share with the students.

RESULTS

Scoring

For each question, students could earn two points for a correct answer, and one point for a partially correct answer. Students earned two points for answers that were correct but misspelled, and for answers that were simple variations in word form from the correct answer (e.g. for the question *The largest group that immigrated to the U.S. during 1846–1860 was from what country?*, students earned two points if they wrote *Irish* instead of *Ireland*). For any answer that required an individual's name, students earned two points for providing the correct last name, and one point for the correct first name without the correct last name. For questions pertaining to general knowledge or terms, students earned two points if their answer provided the correct meaning without necessarily providing the same term as in the expected answer (e.g. for the question *People who worked for women's rights, seeking to improve women's lives and win equal rights, were called what?*, students earned two points if they wrote *suffragist* or *feminist* instead of *American feminist*), or if their answers contained some correct information but were less detailed than a two-point answer. One question inquired about a percentage value (*What percent of plantation owners owned more than half of the slaves in the south?*), and here students earned two points if their answer was within two percentage points of the correct answer (12%), and only one point if their answer was within four percentage points. Finally, for any question pertaining to a date, amendment number or U.S. state, students earned two points by providing the exact correct answer, and no partial credit was given.

Two independent raters utilized this scoring system to score the data from 13 subjects (about 21% of the entire sample) chosen at random. In blind fashion, each rater scored the

DOI: 10.1002/acp

answers provided by students on the final test, along with the 15 items from the Test/Study condition that appeared on the review. Although students were instructed to score their own review sheets for the 15 Test/Study items, the two raters also scored these items in order to verify students' scoring and award partial credit where necessary. Across the three final test conditions and the review session, the points awarded by the two raters were correlated at 97% or higher (all $ps < .001$). Given the high inter-rater agreement, the data from the remaining students were scored in blind fashion by only one of these raters.

Review session

As expected, the Immediate Review Group earned more points on the 15 Test/Study review questions than did the Delayed Review Group. The average number of points earned by the Immediate Review Group was 15.21 (SD = 8.97), whereas the average number of points earned by the Delayed Review Group was 10.61 (SD = 6.88). An independent samples *t*-test confirmed that this difference was significant, $t(60) = 2.28$, $p < .03$, $d = .58$.

Final test

We calculated the number of points earned on the final test for both Immediate and Delayed Groups, and for items that received review through Test/Study, Study or No Review. These scores were analysed using a 2×3 (Group: Immediate vs. Delayed $\times$ Review Method: Test/Study vs. Study vs. No Review) mixed Analysis of Variance (ANOVA). The main effect of review method was significant, $F(2, 120) = 7.87$, $p < .01$, MSE = 4.32. *Post-hoc* tests using Bonferroni's correction revealed that the number of points earned on the final test was significantly higher for items that had been reviewed through Test/Study ($M = 3.08$, SD = 2.74) compared to Study ($M = 2.18$, SD = 2.70), $t(61) = 2.54$, $p < .05$, $d = .32$, and for Test/Study compared to No Review ($M = 1.58$, SD = 1.94), $t(61) = 3.96$, $p < .01$, $d = .51$.[1] Performance did not differ significantly for items that had been reviewed through Study compared to items that received No Review, $t(61) = 1.57$, $p = .41$. The main effect of Group approached significance, $F(1, 60) = 3.67$, $p = .06$, MSE = 9.56, $d = .49$. Out of all 45 items combined, the average number of points earned by the Immediate Review Group was 5.45 (SD = 5.17). The Delayed Group, on the other hand, earned an average of 8.06 points (SD = 5.51). The interaction between group and review method was not significant, $F < 1$.[2] The mean number of points earned across all conditions for both groups are displayed in Figure 1.

The degree of forgetting of course material that had occurred by 36 weeks was, perhaps not surprisingly, very large (fewer than three points, on average, compared to about 15 points at 1 week, and more than 10 points at 16 weeks). For the No Review items, the

[1]Because these comparisons are within-subjects, the effect sizes are corrected for dependence between responses using Morris and DeShon's (2002) Equation (8).

[2]We explored these same effects separately for one-point responses and two-point responses. Considering only one-point responses, there were no main effects for either review method or group, and no interaction (all $Fs < 1.5$). Interestingly, however, the data for the two-point responses revealed a similar pattern to that of the overall data. There was a significant main effect of review method, $F(2, 120) = 7.41$, $p < .01$, MSE = 4.16. *Post-hoc* tests using Bonferroni's correction revealed that retention was significantly better for Test/Study ($M = 2.81$, SD = 2.75) than for Study ($M = 1.87$, SD = 2.53), $t(61) = 2.68$, $p < .05$, $d = .34$,[1] and for Test/Study compared to No Review ($M = 1.39$, SD = 1.87), $t(61) = 3.65$, $p < .01$, $d = .47$. There was no significant difference between Study and No Review, $t(61) = 1.35$, $p = .60$. The main effect of group approached significance, $F(1, 60) = 3.15$, $p = .08$, MSE = 8.82. Out of all 45 items combined, the average number of points earned by the Delayed Review Group ($M = 7.15$, SD = 5.29) was greater than the number of points earned by the Immediate Review Group ($M = 4.83$, SD = 4.97), $d = .45$. The interaction between group and review method was not significant, $F < 1$.

DOI: 10.1002/acp

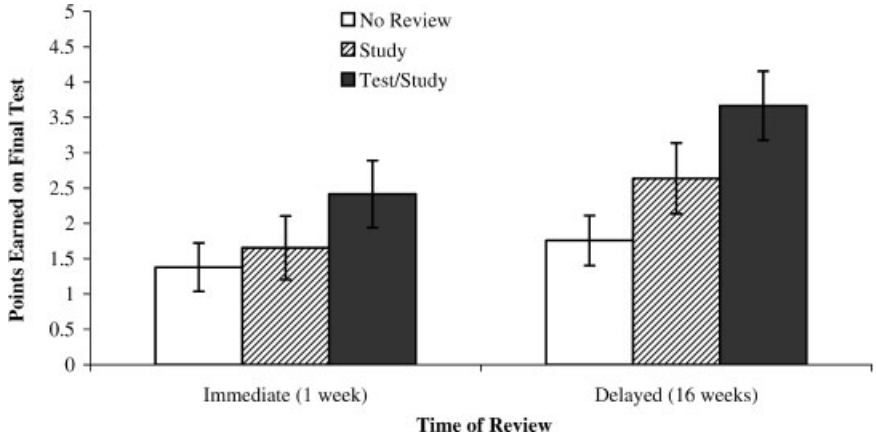

Figure 1. Students reviewed some U.S. history facts through tests with feedback (Test/Study), some through additional study (Study), and some were not reviewed at all (No Review). On a final test 9 months later, students who reviewed the information after a 16-week delay (Delayed Review Group) remembered slightly more than students who reviewed it after a 1-week delay (Immediate Review Group), and retention was significantly better for information reviewed through Test/Study compared to Study and No Review. Error bars represent standard errors

number of points earned on the final test was similar for the Immediate Review Group ($M = 1.38$, $SD = 1.84$) and the Delayed Review Group ($M = 1.76$, $SD = 2.03$), $t(60) = 0.76$, $p = .45$, suggesting that baseline performance was similar across these groups.

GENERAL DISCUSSION

We found that testing was significantly beneficial to 8th grade students' retention of facts from their U.S. history course. These results replicate a number of prior studies demonstrating beneficial effects of testing in laboratory contexts (e.g. Carpenter & DeLosh, 2005, 2006; Carpenter & Pashler, 2007; Carpenter et al., 2006; Kuo & Hirshman, 1996; Roediger & Karpicke, 2006a), in environments that simulate a classroom setting (e.g. Butler & Roediger, 2007), in online college courses (McDaniel, Anderson et al., 2007) and in actual classroom environments with college student learners (e.g. Bangert-Drowns et al., 1991).

Delaying the review session by 16 weeks did not appear to be harmful to final retention. In fact, the Delayed Review Group performed better on the final test than did the Immediate Review group, though this difference did not reach significance. These results are consistent with previous laboratory studies demonstrating beneficial effects of delaying an initial test (Izawa, 1992; Modigliani, 1980; Whitten & R. A. Bjork, 1977). Also consistent with this finding is the well-known tendency for information to be better retained when it has been reviewed on two or more occasions that are distributed across time rather than massed together—commonly referred to as the *spacing effect* or *distributed practice effect* (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1988; Donovan & Radosevich, 1999). Although there have been a number of studies looking at spacing effects in the

DOI: 10.1002/acp

classroom (e.g. Fishman, Keller, & Atkinson, 1968), these have used much shorter spacing and retention intervals than would be of interest in real educational contexts.

Other studies conducted in the laboratory have reported combined benefits of testing and spacing such that the best overall retention occurs for material that has been tested after spaced time intervals (e.g. R. A. Bjork, 1988; Carpenter & DeLosh, 2005; Cull, 2000; Cull, Shaughnessy, & Zechmeister, 1996; Izawa, 1992; Karpicke & Roediger, 2007a; Landauer & R. A. Bjork, 1978; Pashler, Zarow, & Triplett, 2003; Rea & Modigliani, 1985). None of these prior studies, however, explored retention over time periods as long as those used in the current study. The current results replicate these prior findings and effectively extend them to a situation involving real classroom learning over substantially longer intervals of time—16 weeks before an initial review, and 9 months before a final test.

Limitations

The absolute levels of retention after 9 months in the present study were very low. Thus, it seems possible that the absolute benefits observed in these data are very small due to the presence of floor effects. The low level of recall is not surprising given the long retention interval and the requirement for recall. Other research from our lab indicates that adult learners typically forget most of the information they have learned after time intervals much shorter than nine months (Carpenter et al., 2008). Given that floor effects typically mask real effects, these effects do not appear to be problematic for the conclusions we are drawing from the current data.

Moreover, massive forgetting of factual information appears to be very common in real-world learning contexts. Indeed, forgetting may play a major role in why American teenagers have such poor recollection of basic history facts (e.g. Dillon, 2008). For example, when asked multiple-choice questions, fewer than half of these survey respondents could correctly identify when the Civil War had occurred, and about 25% placed the voyage of Columbus after the year 1750, rather than in 1492. It seems highly possible that such information may have been learned at one time, and in the absence of regular review, was subject to forgetting. These examples appear quite relevant to the current finding that much of what students learned about U.S. history was forgotten over the 9-month retention interval. The current results suggest that testing may be one mechanism that has the potential to alleviate some of the negative effects of forgetting over time (e.g. see also Carpenter et al., 2008), and future research would benefit from more investigations of testing in real-world contexts over lengthy time intervals.

The focus of the present report has been on the benefits of testing over re-reading. In the current study, this factor was manipulated within subjects and so these conclusions are rigorously supported. Our results also suggest the existence of a spacing effect (improved retention with a 16-week delay over a 1-week delay). While this effect seems to be both credible and in line with previous research, it should be noted again that students were not individually randomized to a spacing condition, but instead were assigned to groups according to their class section. While potential effects of instructor associated with class section were counterbalanced, we cannot firmly rule out the possibility that some other extraneous factor, which may covary with class section, might have had some influence on the spacing effect.

One very important aspect on which our final test differs from typical classroom exams is the fact that, in the current study, students may not have expected a final test over the material. If they had been expecting a final test—as is typically the case with course

DOI: 10.1002/acp

exams—students would have had opportunities to prepare and so final test scores would have likely been substantially higher. The overall final test scores in the current study, therefore, are not expected to be as high as would be the scores on routine course exams in which students receive grades. In the current study, the opportunity to prepare for the final test was purposefully avoided in the interests of measuring retention under conditions in which additional studying of the material was very unlikely. Compared to a routine course exam, therefore, it is possible that students were not as motivated to do well on the experimental test because of the unexpected nature of the test and the fact that their score did not count towards their course grade. These constraints, however, made it possible to achieve a reasonable degree of control over the amount of exposure to the material.

The questions created for the study were highly factual, often asking about names, dates and general terms. These questions, which were created by the researchers, may have differed from the type of questions that assess knowledge of broader conceptual themes (e.g. how did the North and South develop differently economically, politically and socially?), which were more often the types of questions encountered in class assignments, exams and discussions. It seems likely that the procedures that enhanced memory for discrete facts would also promote retention of general organizing principles (causal structures, hierarchical relationships, etc.), but this issue needs more investigation. Had we assessed 'gist' information rather than specific facts, it is likely that overall performance on the final test would have been substantially greater. Future research would benefit from further exploration of the effects of testing and spacing on direct retention of facts, as well as transfer and generalizability of more conceptual types of knowledge that students learn in their courses.

Implications for learning and instruction

Subject to the limitations just noted, the results of the current study have direct implications for optimizing methods of learning and instruction. We found that a test with feedback clearly improved students' retention more than a restudy opportunity. When students are asked about their study strategies, they generally report re-reading information over and over again in order to commit it to memory (e.g. Carrier, 2003). Results of the current study suggest that this method, despite its popularity, may be highly ineffective. Our study strongly suggests that students would be better off engaging in modes of study that involve testing and re-generation of material, rather than repeated re-reading (e.g. McDaniel, Anderson et al., 2007; Roediger & Karpicke, 2006b). For example, instead of repeatedly reading one's class notes or textbook chapters, students may profit from doing some initial reading and then trying to recall the information they have read. Techniques such as using flashcards and answering textbook review questions may also benefit retention far more than re-reading.

Furthermore, instructors may be well advised to embed questions (rather than just presenting information) during class discussions and assignments, as this would seem to encourage students to recall information. End-of-class quizzes are another means of encouraging students to recall their course material, and there is evidence that such quizzes can be effective (e.g. see Roediger & Karpicke, 2006b).

The power of tests to enhance retention is even greater when students practice recalling the same information more than once (e.g. Allen et al., 1969; Carpenter et al., 2008; Kuo & Hirshman, 1996; Karpicke & Roediger, 2007b). It would seem, therefore, that students in

DOI: 10.1002/acp

classroom contexts would benefit from instructional methods that encourage repeated opportunities to recall the same information. Cumulative exams, which involve repeated testing over the same material, may be quite useful in promoting long-term retention of course material. Future research should explore these potential classroom tools and other methods that encourage the use of recall as a strategy to promote retention of course material.

ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H040108 to the University of California, San Diego. We thank the student participants who volunteered for this study, and we are most grateful for the assistance of the principal and teachers. We thank Lora Chung for attending each class, gathering the course materials and creating the test questions, and Noriko Coburn for her assistance with scoring. We also thank Irene Lai and Melanie Wu for their assistance in assembling the review and test sheets.

REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (in press). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology.* DOI: 10.1002/acp.1391

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *8*, 463–470. DOI: 10.1016/S0022-5371(69)80090-3

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 397–401). New York: Academic Press.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527. DOI: 10.1080/09541440701326097

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619–636. DOI: 10.1002/acp.1101

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, *34*, 268–276.

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

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin and Review*, *14*, 474–478.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition*, *36*, 438–448. DOI: 10.3758/MC.36.2.438

Carrier, M. L. (2003). College students' choice of study strategies. *Perceptual and Motor Skills*, *96*, 54–59. DOI: 10.2466/PMS.96.1.54-56

Carrier, M. L., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, *20*, 633–642.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. DOI: 10.1037/0033-2909.132.3.354

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571. DOI: 10.1037/0096-3445.135.4.553

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235. DOI: 10.1002/(SICI)1099.-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding the understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365–378. DOI: 10.1037/1076-898X.2.4.365

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627–634. DOI: 10.1037/0003-066X.43.8.627

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In R. Bjork, & E. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 317–344). San Diego, CA: Academic Press.

Dillon, S. (February 2008). Survey finds teenagers ignorant on basic history and literature questions. *The New York Times*, A16.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*, 795–805. DOI: 10.1037/0021-9010.84.5.795

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Journal of Educational Psychology*, *59*, 290–296. DOI: 10.1037/h0020055

Izawa, C. (1992). Test trial contributions to optimization of learning processes: Study/test trials interactions. In A. F. Healy, & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes: From learning theory to connectionist theory* (Vol. 1, pp. 1–33). Hillsdale, NJ: Erlbaum.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558. DOI: 10.1080/09541440601056620

Karpicke, J. D., & Roediger, H. L., III. (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. DOI: 10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L., III. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. DOI: 10.1016/j.jml.2006.09.004

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin and Review*, *14*, 219–224.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, *109*, 451–464. DOI: 10.2307/1423016

Kuo, T., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language*, *36*, 188–201. DOI: 10.1006/jmla.1996.2486

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). New York: Academic Press.

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

McDaniel, M. A. (2007). Applying cognition to education: Editorial. *Psychonomic Bulletin and Review*, *14*, 185–186.

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. DOI: 10.1080/09541440701326154

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385. DOI: 10.1037/0278-7393.11.2.371

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review*, *14*, 200–206.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin and Review*, *14*, 225–229.

Modigliani, V. (1980). Immediate rehearsal and initial retention interval in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 241–253. DOI: 10.1037/0278-7393.6.3.241

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125. DOI: 10.1037/1082-989X.7.1.105

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin and Review*, *14*, 187–193.

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. DOI: 10.1037/0278-7393.29.6.1051

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spellings lists. *Human Learning: Journal of Practical Research and Applications*, *4*, 11–18.

Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. DOI: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. DOI: 10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159. DOI: 10.1037/0278-7393.31.5.1155

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology*, *31*, 665–676. DOI: 10.1037/h0054178

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656. DOI: 10.1037/h0063404

Viadero, D. (30 August 2006). Cognition studies offer insights on academic tactics: U.S.-funded projects eye ways of helping students remember more material. *Education Week*, *26*, 12–13.

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*, 240–245. DOI: 10.1111/j.1467-9280.1992.tb00036.x

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 465–478. DOI: 10.1016/S0022-5371(77)80040-6