EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY
2007, 19 (4/5), 494–513

Psychology Press
Taylor & Francis Group

Testing the testing effect in the classroom

Mark A. McDaniel
Washington University in St Louis, St Louis, MO, USA

Janis L. Anderson
Harvard Medical School and Brigham & Women's Hospital, Boston, MA, USA

Mary H. Derbish
Washington University in St Louis, St Louis, MO, USA

Nova Morrisette
University of New Mexico, Albuquerque, NM, USA

Laboratory studies show that taking a test on studied material promotes subsequent learning and retention of that material on a final test (termed the testing effect). Educational research has virtually ignored testing as a technique to improve classroom learning. We investigated the testing effect in a college course. Students took weekly quizzes followed by multiple choice criterial tests (unit tests and a cumulative final). Weekly quizzes included multiple choice or short answer questions, after which feedback was provided. As an exposure control, in some weeks students were presented target material for additional reading. Quizzing, but not additional reading, improved performance on the criterial tests relative to material not targeted by quizzes. Further, short answer quizzes produced more robust benefits than multiple choice quizzes. This pattern converges with laboratory findings showing that recall tests are more beneficial than recognition tests for subsequent memory performance. We conclude that in the classroom testing can be used to promote learning, not just to evaluate learning.

Correspondence should be addressed to Mark A. McDaniel, Department of Psychology, Campus Box 1125, Washington University in St Louis, St Louis, MO 63130, USA. E-mail: mmcdanie@artsci.wustl.edu

This research was supported by Institute of Educational Sciences Grant R305H030339. Mary Derbish's participation was supported by a Collaborative Activity Grant from the James S. McDonnell Foundation. This experiment was presented in part at the 46th Annual Meeting of the Psychonomic Society, Toronto, Canada, November 2005 and at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 2006. We thank Roddy Roediger and Chuck Weaver for helpful comments on an earlier version of this paper and Austin McDaniel, Jesse McDaniel, and Rebecca Roediger for assistance with aspects of the data scoring.

http://www.psypress.com/ecp DOI: 10.1080/09541440701326154

In educational practice, as well as in the educational research literature, testing has been primarily considered an evaluative instrument. However, many researchers who study memory have considered testing from the perspective of its mnemonic benefits. Experimental reports have repeatedly demonstrated that taking a test on studied material promotes subsequent learning and retention of that material on a final test (e.g., Bartlett, 1977; Darley & Murdock, 1971; Hanawalt & Tarr, 1961; Hogan & Kintsch, 1971; Masson & McDaniel, 1981; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985; Whitten & Bjork, 1977). For purposes of exposition we will refer to the memory gains produced by intervening tests as the *testing effect*. Experimental memory research has established that the testing effect is robust across materials and types of tests. Testing effects are observed with word lists (Hogan & Kintsch, 1971; McDaniel & Masson, 1985), paired associate lists (Allen, Mahler, & Estes, 1969; Carrier & Pashler, 1992), pictures (Wheeler & Roediger, 1992), and prose material (Glover, 1989; Roediger & Karpicke, 2006b). Testing effects surface when the intervening tests are different from the final tests: intervening recall tests improve subsequent recognition (Glover, 1989; Lockhart, 1975; Wenger, Thompson & Bartling, 1980) and intervening recognition tests improve subsequent recall (Runquist, 1983). Finally, taking a test is almost always a more potent learning device than additional study of the target material (see Carrier & Pashler, 1992, for recent experimental tests, and Roediger & Karpicke, 2006a, for a review).

Despite this impressive body of evidence, the implications of the testing effect literature for educational practice have been virtually ignored by the educational community and educational research. Echoing this observation, an educationally relevant study on the testing effect was entitled "The 'testing' phenomenon: Not gone but nearly forgotten" (Glover, 1989). Yet, paralleling the basic memory findings, the few studies in the educational literature that have examined the testing effect have found positive benefits of intervening tests on final test performance (Glover, 1989; Spitzer, 1939). Despite these findings, current texts on learning and instruction fail to mention the possible merits of using tests to potentiate learning and retention (e.g., Mayer, 2003; see also Baine, 1986). This omission may be warranted because even the studies appearing in the educational psychology literature have not demonstrated the benefits of testing on material being taught and learned in the classroom.

To fill this critical gap, the purpose of the present research was to experimentally examine the testing effect for content presented throughout the semester in a college course. We were interested in several overarching issues. First, would positive testing effects emerge in the context of a standard course? Testing in a course diverges in important ways from the

implementation of testing in laboratory studies. In a class, there is presumably great variability across students in the amount of studying of the target material and in the interval between study and the intervening testing (testing prior to the criterial tests). In contrast, both of these variables are carefully controlled or manipulated in laboratory studies. Further, the delays between intervening test (i.e., quizzes) and final criterial tests in the classroom can be on the order of days, weeks, or even months. In the laboratory long retention intervals are typically 1 or 2 days (Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Masson & McDaniel, 1981; McDaniel & Masson, 1985); more often these intervals have been on the order of minutes or hours (e.g., Bartlett, 1977). In only very few experiments have benefits of testing in laboratory studies been examined at 1-week or longer intervals (see Roediger & Karpicke, 2006b, and Wheeler, Ewers, & Buonanno, 2003, for 1-week delay studies, and Butler & Roediger, 2007, this issue, for a 1-month delay). Because of these differences between the class environment and the parameters of the laboratory research, it is not certain that the testing effect will generalise to the more variable environment of an actual class.

For the present study, we identified several key issues from the existing literature that could be particularly important for motivating and implementing testing as a learning tool in the classroom. We next briefly review these issues and related findings in developing the rationale for the design of the current experiment.

EXTENDING TESTING EFFECTS TO THE CLASSROOM

One central issue raised by the testing effect findings is the extent to which the repeated exposure of the material stimulated by tests plays a role in the positive impact of intervening tests on final test performance. Some prominent studies in the educational literature do not clarify this issue because conditions were not included that receive extra study instead of intervening tests (e.g., Glover, 1989; Spitzer, 1939). Findings from the basic experimental literature do suggest that testing produces learning/retention advantages beyond that enjoyed from repeated study (provided that the retention intervals between intervening and final testing are greater than several minutes; cf. Roediger & Karpicke, 2006b; Wheeler et al., 2003). For instance, immediate retrieval of once-studied target items benefits performance on subsequent tests more so than does another study presentation of the target material (Hanawalt & Tarr, 1961; Hogan & Kintsch, 1971; McDaniel & Masson, 1985).

To assess the degree to which testing effects in the classroom (if found) reflect mnemonic processes that are more than just additional exposure of

the content, in the current experiment we included an exposure-only condition in which the target facts were presented for reading. In this read only condition, participants were presented with the same information that was quizzed in other conditions. In addition, a control was implemented in which some facts in the course were neither presented for additional reading nor testing. Based on the literature cited, we expected testing effects to emerge (quizzes with feedback would produce better performance on final tests than not tested/read facts), and we expected that testing (quizzing with feedback) would be superior to the read only presentation in terms of boosting performance on final testing.

Another issue addressed in the basic memory literature is the relative benefit of cued recall tests over recognition (e.g., multiple choice) tests. Studies with simple laboratory materials (word or paired associate lists) have found that retrieval through recall benefits subsequent test performance more so than retrieval processes associated with recognition (Cooper & Monk, 1976; Darley & Murdock, 1971; Mandler & Rabinowitz, 1981; McDaniel & Masson, 1985; Wenger et al., 1980; see Glover, 1989, for an identical pattern using short texts as the target material). However, there is at present no published work that contrasts testing effects with the types of quizzes commonly found in a classroom (short answer vs. multiple choice) in the context of an actual course with normal classroom content. To address this unexplored issue, in the present study we manipulated the type of quiz in a college Brain and Behavior course at the University of New Mexico. For target facts, we included quizzes with feedback (see Kang, McDermott, & Roediger, 2007 this issue; McDaniel & Fisher, 1991; Pashler, Cepeda, Wixted, & Rohrer, 2005; and Wininger, 2005, for mnemonic advantages of providing feedback after testing) that either required recall (short answer tests) or recognition (multiple choice tests). Generalising from the findings in the experimental memory literature, we predicted that short answer quizzes would produce greater gains in performance on unit exams than would multiple choice quizzes.[1]

An alternative outcome might be possible as well. For the final criterial tasks (unit exams and a cumulative final) we used multiple choice tests, reflecting the kind of assessment test used in most large college classes. With final multiple choice tests, dynamics of transfer appropriate processing may trump the mnemonic benefits of recall over recognition. Transfer appro-

[1] Our prediction is based on the observation that the multiple choice questions used herein were worded very similarly or in many cases identically (as for the example quiz item in the Appendix) to the factual statements presented in the textbook. Thus, though multiple choice testing on course content is not necessarily identical to laboratory recognition tests, for the present materials we assume that recognition of the target fact would underlie, at least somewhat, performance on the multiple choice tests.

priate processing refers to increased memory performances when prior processing matches processing required for a subsequent test (see Thomas & McDaniel, in press, for an educationally relevant example, and McDaniel, Friedman, & Bourne, 1978; Morris, Bransford, & Franks, 1977; Roediger & Blaxton, 1987, for basic memory results). On this principle, multiple choice quizzes would presumably promote better transfer to the final multiple choice tests than would the short answer quizzes. (This prediction assumes that recognition processes transfer to subsequent recognition more so than recall processes transfer to subsequent recognition.)

Finally, the testing effect previously reported for educationally relevant materials may represent a somewhat brittle effect that is limited to final criterial questions that are identical to those presented in the initial tests (e.g., Spitzer, 1939, used the same question stems across repeated tests, as did Glover, 1989). The testing effect would be optimally valuable in the classroom if it produced learning of a complex fact, rather than learning of a particular answer when given a particular question. Further, in classroom applications, some instructors are understandably reluctant to give identical questions on quizzes and final assessments. Therefore, in the current study we examined the testing effect in a more challenging setting in which the wording of each question was changed between the initial quiz and the subsequent tests (see Materials section).

METHOD

Participants and design

The participants were 35 male and female students enrolled in a web-based Brain and Behavior course at the University of New Mexico who participated voluntarily for extra credit. All participants completed weekly quizzes, two unit tests, and a final exam that were constructed for the experiment; these tests were not used for evaluation in the course. One participant dropped out of the study before Unit 2; his/her data were not included in any of the analyses. Two additional participants dropped out before the final exam; their data were included in the analyses of quiz and unit test performance, but not for the final exam.

The experiment was a 3×2 within-subjects design, with the quiz type (multiple choice, short answer, read only) and target fact exposure (exposed with quiz/reading, not exposed). As detailed below, a set of not-exposed facts was paired with each exposure condition, yielding a complete factorial design. For purposes of exposition, we have labelled the initial variable "quiz type", though note that one level of the variable (read only) is not quizzed but involves only reading the facts.

Materials

Quizzes. Two sets of ten facts were extracted from the assigned material that students were instructed to read each week. Weekly reading assignments consisted of approximately 40 pages from an undergraduate textbook (Rosenzweig, Breedlove, & Watson, 2004). Facts selected for this experiment were included in the textbook but were not those that were emphasised in the course itself. For each of the 10 facts in one set, a second fact was taken from the same paragraph to create the second set of facts. For example, the two facts below were taken from the same paragraph of the material and would be assigned to different sets for counterbalancing purposes:

> Set 1: All preganglionic axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter.
> Set 2: Parasympathetic postganglionic axons release acetylcholine as a neurotransmitter

During each of 6 weeks one set of facts was quizzed/read and one set was not quizzed/not read. These sets were counterbalanced across participants. In addition, each week across participants each set of facts was presented in one of three "quiz" forms: multiple choice (MC), short answer (SA), or read only (RO). Combining these two counterbalancing factors yielded six counterbalancing groups to which the participants were randomly assigned. Table 1 provides a description of this counterbalancing design and the number of students assigned to each group (see the Appendix for an example of each of the three quiz forms). The experiment spanned 6 weeks of the course, thereby allowing two replications of the design per student.

TABLE 1
Quiz counterbalancing design

Quiz type	*Counterbalancing condition*	*Fact set*		*N*
		A	*B*	
RO	1	Read	Not read	6
	2	Not read	Read	4
MC	1	Quizzed	Not quizzed	5
	2	Not quizzed	Quizzed	7
SA	1	Quizzed	Not quizzed	7
	2	Not quizzed	Quizzed	5

Counterbalancing conditions have unequal *N* because some students dropped the course after the initial random assignments to the counterbalancing conditions. RO = read only; MC = multiple choice; SA = short answer.

Feedback. Feedback was constructed for each of the weekly quizzes. Feedback included the test questions, question number (or the read fact), the correct answer, and the participant's response. For the multiple choice questions, all answer choices were given as well. For short answer questions, the question was displayed along with the participant's response and all correct answers. Finally, for read only facts the feedback would always display the fact that was presented on the quiz and a participant response of "I have read the above statement". Examples of all three types of feedback are provided in the Appendix.

Unit tests. Two 60-item multiple choice unit tests were constructed, one for the first 3 weeks of facts presented in the assigned readings and another for the second 3 weeks of facts. The 60 items comprised the entire set of facts quizzed during the previous 3-week period. Note, however, for any particular participant, 30 items had been quizzed (10 MC, 10 SA, 10 RO) and 30 items had not been quizzed (10 yoked to each of the MC, SA, and RO conditions). Feedback was not provided for the unit exams.

To test for retention of the complete conceptual relation (rather than retention of a particular answer provided in a quiz), each fact was tested such that the answer required for a quiz item was now embedded in the question stem, and an alternative portion of the fact was required for the answer. As an example, the quiz wording and unit-test wording for a fact would read as follows:

Quiz wording: All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter:

a. acetylcholine
b. epinephrine
c. norepinephrine
d. adenosine

Unit-test wording: All _____ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter.

a. preganglionic
b. ionotropic
c. hypothalamic
d. adenosine

Final exam. A multiple choice final exam was constructed. The final exam consisted of all 60 items from both unit tests for a total of 120 items. Half of the items were presented in the same wording as the quiz and half were presented in the same wording as the unit test. Therefore for the final exam, students had seen the exact wording of the question previously.

Procedure

Each week participants were assigned approximately 40 pages of textbook reading in the course. As participants in the research, they were instructed to log on at the end of each week and complete the 10-item quiz over that week's readings. Each week, any particular participant received his or her quiz in a different test format (MC, SA, or RO). On the week when the participant received the RO condition, they simply read the designated target facts and clicked a button for the response "I have read the above statement." Participants were allowed 10 min to complete each quiz; immediately after finishing they were provided access to feedback. The participant clicked on "submit quiz" and received a confirmation statement that their quiz was successfully submitted along with a "view results" link that took them to the results display. The participants could inspect the feedback for as long as they wanted and as many times as they wanted within a week of completing the quiz.

After 3 weeks of quizzes (one MC, one SA, and one RO) the participants were instructed to take the first unit test, with all participants receiving the same unit test. Next, participants were given another 3 weeks of quizzes. Similar to the first 3 weeks, participants were given their quiz in a different test format each week and were provided feedback after each quiz. After completing the second set of three quizzes, participants were instructed to take the second unit test, which tested only the material presented in the second 3 weeks of quizzes. Several weeks after completing the second unit test, participants were instructed to take the final cumulative exam. Students were told that this was a practice cumulative exam that might help them on the in-class final. To avoid contamination, none of the facts tested in the experiment was tested on the actual course exams.[2]

RESULTS AND DISCUSSION

Quiz performance

The mean proportions of quiz questions answered correctly for units one and two are shown in Table 2. A 2×2 within-subjects analysis of variance (ANOVA), with the factors of quiz type (MC or SA) and unit (1 or 2), indicated that there was a main effect of question type such that participants were more likely to answer MC questions correctly than SA questions,

[2] Actual course examinations could not be used in the experiment because the Institutional Review Board would not allow the experiment to be conducted as a required part of the course. Consequently, using the material tested in the course as target material for the experiment was judged as possibly coercive and therefore inappropriate.

TABLE 2
Proportion correct on quizzes

	Unit 1	*Unit 2*
MC	.49 (.21)	.37 (.22)
SA	.17 (.20)	.21 (.20)

Standard deviations are in parentheses. MC = multiple choice; SA = short answer.

$F(1, 33) = 54.12$, $MSE = 0.04$ (for all analyses the alpha level for determining significance was set at .05). There was also a significant interaction between type and unit such that the benefit of answering MC questions compared to SA questions was greater in Unit 1 than in Unit 2, $F(1, 33) = 4.50$, $MSE = 0.04$. The advantage of multiple choice performance over short answer is entirely consistent with the principle that recognition is a less demanding retrieval task than recall. We next examine the extent to which these initial tests influenced performance on subsequent criterial tests.

Unit test performance

The mean proportions of questions answered correctly for the unit one and two multiple choice tests are shown in Table 3. A $2 \times 3 \times 2$ within-subjects ANOVA, with the factors of unit (1 or 2), test type (MC, SA, or RO), and preexposure (quizzed or not quizzed) was conducted on these data. Importantly, there was a main effect of preexposure such that performance was generally better when facts were previously quizzed than when the facts were not quizzed, $F(1, 33) = 14.77$, $MSE = 0.03$. To inform the issues outlined in the introduction, planned contrasts between quizzed and not quizzed items for each quiz type were computed (collapsed across unit; see Figure 1). The advantage of quizzed over not quizzed facts was significant for multiple choice quizzes, $F(1, 64) = 4.00$, $MSE = 0.03$, and for the short answer quizzes, $F(1, 64) = 16.00$, $MSE = 0.03$. There was no

TABLE 3
Proportion of questions answered correctly on Unit 1 and Unit 2 tests

	Unit 1 (N = 34)		*Unit 2 (N = 34)*	
	Quizzed	*Not quizzed*	*Quizzed*	*Not quizzed*
MC	.55 (.20)	.50 (.22)	.44 (.25)	.37 (.21)
SA	.61 (.24)	.49 (.21)	.53 (.24)	.42 (.22)
RO	.51 (.22)	.51 (.21)	.43 (.24)	.38 (.19)

Standard deviations are in parentheses. *N* refers to number of participants. MC = multiple choice; SA = short answer; RO = read only.

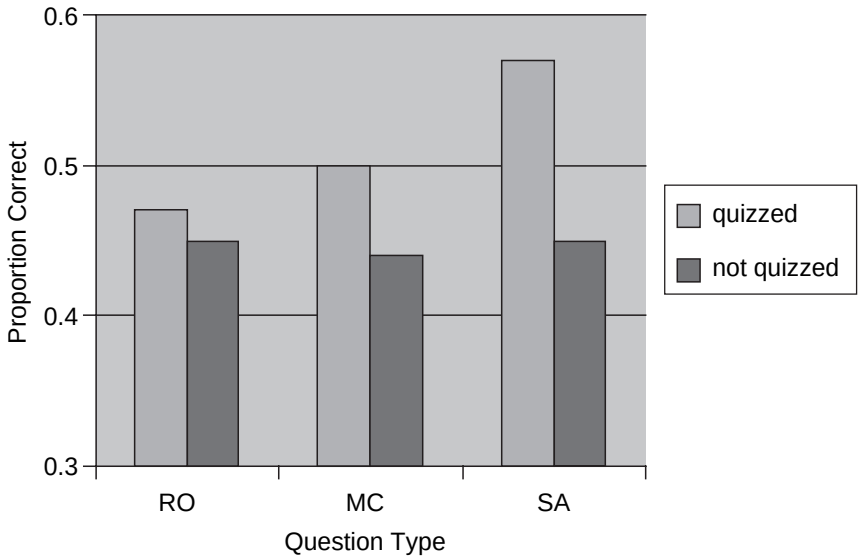

Figure 1. Unit exam performance of quizzed versus not quizzed items collapsed across units.

significant advantage of presenting facts for reading relative to not presenting the facts ($F < 1$). These patterns clearly reveal a learning benefit of prior quizzing with feedback—a testing effect. Further, this testing effect cannot be interpreted as a mere exposure effect because exposure per se of the facts (the read only condition) produced no significant benefit on the unit test.

There was also a main effect of quiz type such that facts assigned to the SA quiz conditions (exposed and nonexposed) were more accurately learned and retained than facts assigned to either the MC or RO questions, $F(2,\ 66) = 3.42,\ MSE = 0.03$. Though, the interaction between the factors of test type and preexposure was only marginally significant, $F(2, 66) = 2.55$, $MSE = 0.03$, $p = .07$, examination of Table 3 suggests that the main effect of quiz type was carried by the quizzed (exposed) facts. Planned comparisons confirmed this impression. For exposed facts, there was a significant advantage of short answer quizzing over multiple choice quizzing, $F(1,\ 66) = 5.44,\ MSE = 0.03$, and read only questions, $F(1,\ 66) = 11.11$, $MSE = 0.03$, but no significant advantage of multiple choice quizzing relative to reading ($F < 1.0$). For facts with no prior exposure, as expected there were no differences as a function of the particular condition to which the fact was assigned (largest $F < 1.0$). Finally, there was a main effect of unit such that participants performed better on the Unit 1 test than they performed on the Unit 2 test, $F(1, 33) = 18.47$, $MSE = 0.06$. The instructor's impression (JLA) was that students may have been spending less time on the

readings midway through the semester (Unit 2 testing) than during the initial part of the semester (Unit 1 testing).

The above results seem consistent with findings in the basic memory literature, using very different materials, showing that recall promotes retrieval processing that is more mnemonically potent than does recognition (Bartlett, 1977; Glover, 1989; McDaniel & Masson, 1985). That is, recall typically produces a greater testing effect than recognition. Indeed, the testing-effect patterns with these complex facts exhibit some similarities to those reported by McDaniel and Masson (1985) using word lists. In that study, cued recall produced significantly better performance on a subsequent cued recall test than did recognition, but importantly half of the time the cues that prompted recall on the final test were different than those that were provided for earlier study and testing. This pattern prompted McDaniel and Masson to suggest that retrieval through recall produces enriched, variable encoding of the target information, more so than retrieval through recognition. The present findings parallel this idea, as cued recall quizzes enhanced performance significantly more than did recognition quizzes on a subsequent test in which the retrieval cues had been altered (i.e., a different question stem was provided than during quizzing).

Interpretation of the potential mechanisms underlying the present effect is more complicated, however. As would be typical for many classroom situations, the present study also provided feedback to students on their quiz responses. Thus, several possible explanations for the over all benefit of SA quizzes over receiving MC quizzes are likely: (a) retrieval failure when recall was attempted (SA) elicited more attentive or effective processing of the feedback than did recognition failure (MC), (b) the SA quiz benefits reflected the potency of recall retrieval relative to recognition retrieval, or (c) both. To gain some insights into these possibilities we conducted a set of conditionalised analyses to attempt to isolate the retrieval and feedback effects of the different quiz types.

Feedback effect. We examined whether the benefits of feedback for missed items were differential across MC and SA quizzes. In addition, we were interested in the general question of whether exposure to the facts as feedback (after missing the fact on a quiz) promoted more learning than exposure to the facts through reading. Accordingly, we calculated the proportion correct on each unit test for items missed on the initial quizzes and compared those values to performance for RO items. A 3 (MC, SA, or RO) $\times$ 2 (Unit 1 or Unit 2) within-subjects ANOVA revealed a main effect of quiz type. Examination of Table 4 shows that facts missed on the SA quiz were more likely to be answered correctly on the unit than facts missed on the MC quiz or simply read, $F(2, 64) = 3.18$, $MSE = 0.04$. Planned contrasts

TABLE 4
Proportion of questions answered correctly on unit after being answered incorrectly on quiz or having been read

	Unit 1 (N = 33)	*Unit 2 (N = 33)*
MC	.54 (.23)	.40 (.26)
SA	.61 (.25)	.47 (.24)
RO	.50 (.22)	.43 (.25)

Standard deviations are in parentheses. *N* refers to number of participants. MC = multiple choice; SA = short answer; RO = read only.

confirmed that missed SA facts were answered correctly (on the unit tests) significantly more often than missed MC facts, $F(1, 64) = 4.90$, $MSE = 0.04$, and significantly more often RO facts, $F(1, 64) = 6.40$, $MSE = 0.04$. The ANOVA also found a main effect of unit such that overall performance was higher for Unit 1 than for Unit 2, $F(1, 32) = 10.77$, $MSE = 0.06$, but there was no interaction of unit with quiz type ($F < 1$).

Note that possible item-selection effects may have contributed to the differential emergence of a positive effect of feedback across missed SA and MC items. Participants missed fewer MC items than SA items, so the missed MC facts were likely relatively harder facts than missed SA facts and certainly harder than the set of entire RO facts. This observation raises two issues. The first is whether positive feedback effects would appear even for missed SA items that were relatively difficult. The second is whether positive feedback effects would appear for missed MC items when a relatively comparable set of RO facts is used as a baseline. In order to examine these issues, we identified a subset of difficult SA questions (questions that were missed on average more than 70% of the time on the initial SA quiz). The resulting subset of difficult SA items consisted of approximately 75% of the original quizzed items. Next we calculated unit test performance sampling only these "difficult" facts, and for each subject only the facts within this set that were answered incorrectly on the quizzes (or all of the difficult facts in the RO condition). The pattern for the "difficult" questions was identical to that found in the initial analysis (thus we dispense with reporting means).

The results are compelling for feedback effects after missing a short answer quiz item. Clearly, learning and retention were better when students were given feedback after missing a short answer question than reading the fact (twice) without being quizzed. In the analyses conducted, the items analysed would overlap considerably across missed SA questions and RO conditions, so that possible item-selection artifacts for this comparison are unlikely. (Item differences would favour the RO condition anyway.) Importantly, the advantage of missed SA items obtained even though the

corrected answer on the quiz was not the response that would be required for correct performance on the unit test. Thus, it appears that giving feedback to items not recalled promoted integrated learning of the elements comprising the tested items.

The findings suggest that feedback for missed multiple choice facts did not benefit learning more so than additional exposure (RO). However, because the analysis with the set of difficult facts still could not perfectly equate the set of items compared across MC (only missed items) and RO (all difficult items) presentations, the possibility remains that feedback after missed multiple choice items could produce more learning than RO exposure.

Retrieval effect. To investigate the contribution of retrieval processing to the testing effects for SA and MC questions, we calculated the probability of a correct response on the unit test conditionalised on correct quiz performance. Doing so yielded a robust advantage for facts recovered via recall (SA) than for recognition (MC), but this result is not overly telling (there were far fewer recalled items than recognised items, likely producing item selection effects). Accordingly, rather than report those raw conditionalised results, we instead report the results of a conditionalised analysis based on a limited set of easier items in order to better equate the item difficulty across SA, MC, and RO conditions. The set of easier items were those answered correctly more than 50% of the time on the multiple choice quiz. The analysis is reported for each unit, to minimise deletion of cases due to missing data (e.g., in the SA condition).

The means are shown in Table 5, along with the number of participants who had a complete set of scores (in general, participants with missing data were those who failed to answer any SA questions on the quiz). A within-subjects ANOVA for Unit 1 (with the factors of test type and exposure) found no significant effect. The ANOVA for Unit 2 found a significant effect of quiz type, $F(2, 34) = 5.08$, $MSE = 0.11$. Contrasts confirmed that recalling an answer (SA retrieval) conferred a robust increase in performance rela-

TABLE 5
Proportion of easiest questions answered correctly on the unit test after being answered correctly on the quiz or having been read

	Unit 1 (N = 13)	*Unit 2 (N = 18)*
MC	.71 (.33)	.58 (.34)
SA	.62 (.46)	.76 (.39)
RO	.55 (.27)	.41 (.34)

Standard deviations are in parentheses. *N* refers to number of participants. MC = multiple choice; SA = short answer; RO = read only.

tive to reading a fact, $F(1, 34) = 10.21$; recognising an answer (MC retrieval) produced only a marginal advantage relative to reading, $F(1, 34) = 2.41$, $p = .13$.

The Unit 2 results support the idea that retrieval of target information benefits retention more than additional study (the RO condition), with recall rather than recognition-like processes producing the retrieval benefit (see McDaniel & Masson, 1985, for similar findings with word list materials). Further, this pattern also suggests that the overall mnemonic benefit of receiving SA quizzes relative to MC quizzes or to additional presentation of target content (RO condition) was in part due to retrieval effects (at least for the section of the course for which students did not fare as well in general). That is, correct retrieval (recall) on short answer questions appeared to potentiate later test performance.

Final exam performance

The mean proportions for final exam performance were submitted to a $2 \times 2 \times 3 \times 2$ ANOVA, with the factors of unit (1 or 2), quiz preexposure (quiz or no quiz), type of quiz (MC, SA, or RO), and wording (same as quiz or same as unit). The patterns of quiz preexposure evident on the unit test mostly persisted to the final cumulative exam. Performance was generally better for facts exposed in the quiz condition than for facts that were not exposed, $F(1, 31) = 6.14$, $MSE = 0.03$. Critically, there was an interaction between quiz preexposure and quiz type such that the benefit of preexposure significantly varied as a function of quiz type, $F(2, 62) = 4.73$, $MSE = 0.04$. The means representing the interaction are shown in Figure 2.

Planned contrasts of quiz preexposure versus no quiz for each quiz type were calculated to more specifically identify the locus of the interaction. The advantage of quizzed over not quizzed facts was significant for SA quizzes, $F(1, 62) = 6.48$, $MSE = 0.04$, and marginally significant for MC quizzes, $F(1, 62) = 2.88$, $MSE = 0.04$, $p = .09$. RO preexposure produced no benefit relative to no preexposure ($F < 1$). The advantage of quizzed SA items over quizzed MC items that emerged on the unit tests did not reach significance for the final examination performance, $F(1, 62) = 2.00$, $MSE = 0.04$. Additionally, the ANOVA showed a main effect of unit such that overall performance was better for Unit 1 items ($M = 0.54$) than Unit 2 items ($M = 0.47$), $F(1, 31) = 18.87$, $MSE = 0.04$. Finally, there was a significant interaction between unit and quiz preexposure such that a benefit of preexposure was obtained for facts from Unit 2 (M quizzed $= 0.50$, nonquizzed $= 0.44$) but not for facts from Unit 1 (M quizzed $= 0.54$, nonquizzed $= 0.54$), $F(1, 31) = 5.08$, $MSE = 0.03$.

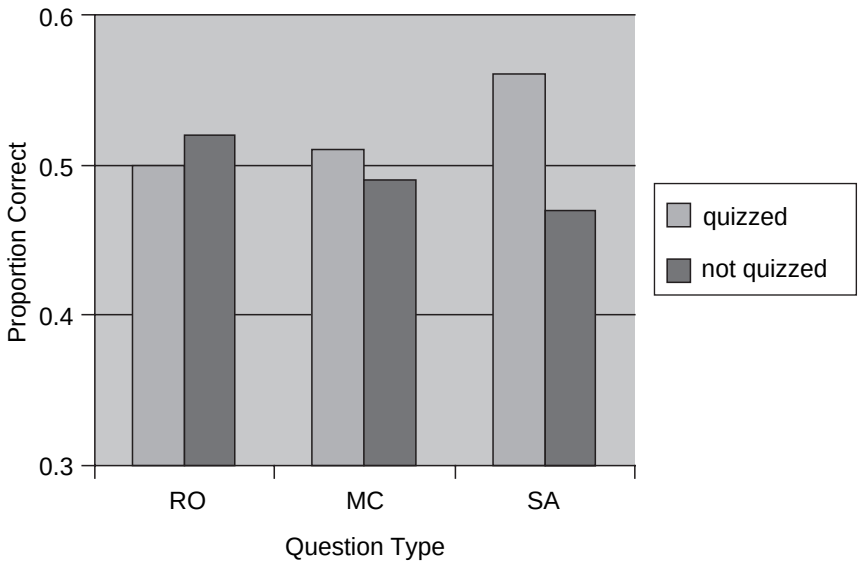

Figure 2. Final exam performance for quizzed versus not quizzed items.

Test spacing

Because the course was web-based and testing was self-initiated, the dates for which participants logged on to the website were monitored. We calculated how many days apart each participant took the quizzes, unit tests, and final. The means were calculated for the number of days between each consecutive pair of tests and are shown in Table 6. As can be seen from the table, students generally adhered to taking the quizzes about 1 week apart, and took the unit test several days after taking the last quiz. On average, the cumulative final exam was taken about 40 days (5 weeks) after the Unit 2 test, though the range varied from 30 to 56 days.

TABLE 6
Average number of days between tests

Q1–Q2	*Q2–Q3*	*Q3–U1*	*Q4–Q5*	*Q5–Q6*	*Q6–U2*	*U2–F*
7.71	7.76	8.87	4.91	11.00	7.58	40.23

Q = weekly quiz, U = unit test, F = final exam.

CONCLUSIONS

Quizzing improved performance on two unit exams and a cumulative final exam for content covered in a college course relative to content that was not quizzed. Consistent with basic research on the testing effect, the benefit for short answer quizzing was more robust than the benefit for multiple choice quizzing. These findings demonstrate that even in the face of the variable conditions found in a course setting, testing enhances learning and retention (e.g., variability in studying and completing assignments across students, variability in motivation, variability in delays between study and quizzing and between quizzing and criterial testing).

Moreover, the present testing effects were obtained despite the question frames changing from the quizzes to the unit tests. This represents a much more demanding transfer task than implemented in some previous testing effect studies using class-related materials (i.e., text or lectures). In previous studies, the question frames have been repeated across initial and final tests (e.g., Butler & Roediger, 2007 this issue; Glover, 1989; Spitzer, 1939). Thus, the present results demonstrate that learning effects from testing extend beyond mere reproduction of previous quiz answers.

Quizzing that required recall of target information (short answer quizzes), but not quizzing that required recognition (multiple choice quizzes), was more effective than presenting the target information for reading. The present conditions were possibly not optimal for producing the most potent testing effects. In the current experiment, facts were quizzed just once; repeated quizzing for target content produces more robust gains on final tests, even compared against repeated study opportunities (e.g., Roediger & Karpicke, 2006b; Wheeler & Roediger, 1992). Also, as noted above the question frames were not the same across quizzes and unit exams. With repeated quizzing or with similar question frames for quizzes and exam, multiple choice quizzing might become more mnemonically potent than additional exposure.

The present quizzing effects likely depended on feedback being provided for quizzed items (Kang et al., 2007 this issue, provide direct evidence on this point in a laboratory experiment). Conditional analyses of performance on missed items showed that the testing effects were obtained for missed items, at least for the SA quizzes. In light of Pashler et al.'s (2005) report that subsequent performance on missed items not given feedback is very poor, it seems likely that the feedback was instrumental in boosting performance for items not correctly answered on the quiz (again, mainly for short answer questions) (see also Wininger, 2005, for positive benefits of providing feedback on quizzes). This finding raises the interesting question for basic research of why processing feedback of a missed item is more effective than exposure to the target content in the absence of a quiz.

There was also evidence that successful retrieval—primarily that required by SA rather than MC questions—contributed to the present testing effect.[3] Though the analysis of conditionalised performances on which this conclusion is based were not entirely consistent (significant retrieval effects were not found for Unit 1), given the present patterns and previous laboratory findings (e.g., Glover, 1989; McDaniel & Masson, 1985), it seems reasonable that both successful retrieval (recall) and processing of feedback after not being able to retrieve the correct answer contributed to the effective use of short answer quizzing in this experiment.

In closing, the fundamental implication of our findings is that testing to enhance learning should be seriously considered in pedagogical theory and practice. There are compelling strengths of an intervention that uses testing to enhance learning. First, test enhanced learning can be implemented for a variety of course contents. Courses that are heavily fact based, in which students are responsible for learning a large body of facts, seem to be especially good candidates for test enhanced learning. Second, application of test enhanced learning to courses at all levels of the curriculum from primary school to college is straightforward. Though little research has been conducted on testing effects with children, the available work indicates large testing effects for children in elementary school (Metcalfe, Kornell, & Son, 2007 this issue, Exps 1 and 2; Spitzer, 1939). Third, implementing test enhanced learning requires no change in curriculum or teaching style. Indeed, for courses in which web-based assistance is possible, using testing as a learning tool would not require valuable class time. Educational theory and practice would do well not to forget the use of testing as a tool to promote learning and retention.

REFERENCES

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*, 463–471.

Baine, D. (1986). *Memory and instruction*. Englewood Cliffs, NJ: Educational Technology.

Bartlett, J. C. (1977). Effects of immediate testing on delayed retrieval: Search and recovery operations with four types of cue. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 719–732.

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.

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

[3] The mnemonic benefits of retrieving an answer for a short answer quiz question may appear related to the mnemonic benefits of generating a target item during study (i.e., the generation effect, Jacoby, 1978; Slamecka & Graf, 1978). In considering this issue, Carrier and Pashler (1992) suggested, however, that the two effects emerge for different reasons.

Cooper, A. J. R., & Monk, A. (1976). Learning for recall and learning for recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 131–156). New York: Wiley.

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*, 66–73.

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

Hanawalt, N. G., & Tarr, A. G. (1961). The effect of recall upon recognition. *Journal of Experimental Psychology*, *62*, 361–367.

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

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

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

Lockhart, R. S. (1975). The facilitation of recognition by recall. *Journal of Verbal Learning and Verbal Behavior*, *14*, 253–258.

Mandler, G., & Rabinowitz, J. C. (1981). Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 79–90.

Masson, M. E. J., & McDaniel, M. A. (1981). The role of organization processes in long-term retention. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 100–110.

Mayer, R. E. (2003). Memory and information processes. In W. M. Reynolds & G. E. Miller (Eds.), *Educational psychology: Vol. 7. Handbook of psychology* (pp. 47–57). Hoboken, NJ: Wiley.

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

McDaniel, M. A., Friedman, A., & Bourne, L. E. (1978). Remembering the levels of information in words. *Memory and Cognition*, *6*, 156–164.

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory and Cognition*, *17*, 423–434.

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

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based program to enhance study efficacy in a high and low-risk setting. *European Journal of Cognitive Psychology*, *19*, 743–768.

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

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8.

Roediger, H. L., III, & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. In D. S. Gorfein & R. R. Hoffman (Eds.), *Memory and learning: The Ebbinghaus Centennial Conference* (pp. 349–379). Hillsdale, NJ: Erlbaum.

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

Rosenzweig, M. R., Breedlove, S. M., & Watson, N. V. (2004). *Biological psychology: An introduction to behavioral and cognitive neuroscience*. Sunderland, MA: Sinauer.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory and Cognition*, *11*, 641–650.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604.

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

Thomas, A. K., & McDaniel, M. A. (in press). The negative cascade of incongruent task-test processing. *Memory & Cognition*.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 545–559.

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

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.

Whitten, W. B., & Bjork, R. A. (1977). *Learning from tests: Effects of spacing: Journal of Verbal Learning and Verbal Behavior*, *16*, 465–478.

Wininger, S. R. (2005). Using your tests to teach: Formative summative assessment. *Teaching of Psychology*, *32*, 164–166.

APPENDIX: SAMPLE QUESTIONS AND FEEDBACK

Questions

(MC) All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter:
- a. acetylcholine
- b. epinephrine
- c. norepinephrine
- d. adenosine

(SA) All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter.

(RO) All preganglionic axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter.

Feedback

(MC) All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter:
- a. acetylcholine
- b. epinephrine
- c. norepinephrine
- d. adenosine

Student Response: b. epinephrine Correct Answer: a. acetylcholine

(SA) Question 1 All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter.

Student Response: acetylcholine Correct Answer: acetylcholine

(RO) Question 1 All preganglionic axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter.

Student Response: I have read the above statement.