Testing Effect in a College Class

Sau Hou Chang

Indiana University Southeast

Completed in

July 2013

Abstract

The present study aimed to investigate the *testing effect* in a regular college class. The research question was whether there were any differences in unit tests performance under different learning conditions. Thirty-three college students at a Midwest university participated in the present study. A repeated measures analysis of variance (ANOVA) was used with the independent variable of learning conditions (study-study, test-test and control). The dependent variables were three unit tests. Results showed that the mean unit test score in the test-test condition (66.29%) was significantly higher than that in the study-study condition (59.47%). However, the mean unit test score in the control condition (61.52%) did not differ from those in the test-test condition and study-study condition. Students rated the pre-/post-tests in the test-test condition (5.9), and the control condition when they were given only lecture notes (5.13) more helpful than the pre-/post-study statements in the study-study condition (3.97).

Keywords: testing effect, college students, educational psychology class

Testing Effect in a College Class

From a meta-analysis of 35 studies of the effects of frequency of classroom testing, Bangert-Drowns, Kulik, and Kulik (1991) found that the use of frequent classroom testing does increase students' achievement performance and improve students' attitude toward instruction. Testing frequency in the control condition was the most important predictor of effect size. Effect sizes were moderately high when the frequently tested group was compared with a control group that received no tests. However, increasing test frequency may regularly improve post-instruction achievement, but the improvement diminishes as test frequency increases. To improve performance, how frequent should tests be given to students in the classrooms?

Fulkerson and Martin (1983) gave one objective 25-item test every two weeks to one group who were allowed to review each completed test and one objective 50-item test identical to the two combined 25-item tests every four weeks to another group who were not given each completed test to review. A 60-item final exam that had not been used on previous exams was given at the end of the semester. They found that the frequent, shorter tests over smaller amounts of material in the experimental group led to significantly better test-by-test performance than the less frequent, longer tests over larger amounts of material in the control group.

The frequency of testing was increased to one test a week in Landrum (2007)'s class. He created a 20-item multiple-choice quiz for each chapter every week of the course. Students completed quizzes in the class and picked up printouts of their quiz performance indicating their answers and the correct answers the following week. At the end of the semester, students completed a cumulative final exam which consisted of half the items previously presented on quizzes. Results showed that there was a significant correlation between average

quiz score and cumulative final exam score. No matter what a student's quiz percentage ranking (top third, middle third, bottom third), student scores increased when comparing the change from overall quiz percentage to overall final exam percentage. In addition, the performance on similar items on the cumulative final was slightly higher than on the original quiz.

Leeming (2002) further increased the frequency of testing to two tests in a week. Students in the test group were given a test during the first 10 to 15 minutes of each class, totaling 22 to 24 tests throughout the semester. Most tests had two short-essay questions taken from the pool of study questions provided in the syllabus, and about five short-answer questions based on material from the text or lecture. After the test, the correct answers were discussed in 2 to 3 minutes. Students in the control group were given four tests throughout the semester. At the end of the semester, all students were given a 2-hour retention test of the first four chapters of the textbook and associated lecture material. The test contained short-essay, multiple-choice, and fill-in-the-blank questions from the class tests previously given to the students. The final course grades and retention test of the test group who had a short test at the start of every class were better than those of the control group who had four tests in the semester.

Not only does frequent testing increase students' achievement, Roediger and Karpicke (2006) demonstrated that testing has a greater positive than studying the material on future retention. They designed three learning conditions: Students studied a short prose passages four times, studied it three times and took one test, or studied it once and took three tests. All students took a final test either 5 minutes or 1 week later. When the final test was given after five minutes, repeated studying improved recall relative to repeated testing. However, on the delayed tests, prior testing produced substantially greater retention than studying, even

though repeated studying increased students' confidence in their ability to remember the material. They suggested that the enhanced performance of testing is a result of the retrieval processes that reactivate and operate on memory traces. Such an improved performance from taking a test is known as *testing effect*.

Einstein, Mullet, and Harrison (2012) found *testing effect* even when the test was taken only once. They asked students to read one short prose passage using a Study-Study strategy and another passage using a Study-Test strategy. For the Study-Study condition, students read one of the passages for a 4-min period and then reread it for a second 4-min period. For the Study-Test condition, students read the other passage for 4 min and wrote as much as they could remember for 4 minutes. During the study period, students could highlight, underline, or take notes when they read the passage. After testing students' memory for both texts with short-answer quizzes, they found that the performance was higher in the Study-Test condition than the Study-Study condition.

To investigate whether the *testing effect* generalizes from the laboratory to the classroom, McDaniel, Anderson, Derbish and Morrisette (2007) redesigned a web-based college course. They had students taken weekly quizzes, two unit tests and a final exam but these tests were not used for evaluation in the course. During each of the six weeks, students were assigned approximately 40 pages of textbook reading. At the end of each week, they completed a 10-item quiz over that week's readings in a different test format (multiple-choice, short answer, read only) and feedback was given following the quiz. In the "read only" condition, students read the designated target facts without taking a test. After three weeks, a unit test in multiple-choice was given of the previously quizzed and read items and the not-previously-tested items, but no feedback was given. Several weeks after completing the second unit test, students took the final exam which consisted of all items from both unit

tests, with half of the items presented in the same wording as the quiz and half presented in the same wording as the unit test. Results showed that quizzing, but not additional reading, improved performance on the unit test and final exam relative to material not targeted by quizzes. Further, short answer quizzes produced more robust benefits than multiple choice quizzes. It showed that recall tests are more beneficial than recognition tests for subsequent memory performance.

Even though McDaniel, et al. (2007) were able to demonstrate the *testing effect*, the class they used was not a typical college class. First, the unit tests and final exam were not used in grading. Second, the unit tests and the final exam were previously quizzed. Third, the web-based course only lasted for six weeks.

The present study aimed to further investigate the *testing effect* in a regular college class. The research question was whether there were any differences in unit test performance under different learning conditions (study, test, control). This study was different from previous study: The unit tests were counted towards students' grades, items in the unit tests were not previously used, and the class was a traditional face-to-face college course lasted for one semester.

Method

Participants

Thirty-three college students (Male =3; Female =30) at a Midwest university participated in the present study in partial fulfillment of an Educational Psychology course requirement. They enrolled in two sections of the course taught by the researcher in the same semester. The procedures met all American Psychological Association (APA) ethical principles for use of human subjects (APA, 2002), and participants were provided informed consent in accordance with guidelines set by the Institutional Review Board of the

university.

Materials

Pre-/post-tests and unit tests.

The Educational Psychology course was divided into three units: learners, learning, and teaching. Each unit had four lessons, and each lesson had subtopics. The number of questions on each subtopic was proportionally constructed. Twenty multiple-choice questions were constructed from the test manuals of the textbook and associated lecture material for each lesson: ten for the testing condition and ten for the unit test. Therefore, there were ten multiple-choice questions for each pre-/post-test which had identical questions, and 40 multiple-choice questions for each unit test which covered four lessons. None of these questions were repeated and only questions for the unit tests were counted towards the grade.

Pre-/post-study statements.

The ten multiple-choice questions for the testing condition from each lesson were rewritten as statements for the study condition. Therefore, there were ten statements for each pre-/post-study which had identical statements. The order of the multiple-choice questions, as well as the options, and study statements were randomized so that no students would receive the questions, the options of the questions, or the study statements in the same order.

Performance and helpfulness scales.

A performance scale was constructed for students to rate how well they would perform at the unit test (1=not well at all, 10=extremely well). A helpfulness scale was constructed for students to rate the helpfulness of the pre-/post-tests, the pre-/post-study statements, or the lecture notes in their preparation for the unit tests (1= not helpful at all, 10=extremely

helpful).

Design

A repeated measures analysis of variance (ANOVA) was used to detect differences in performance under different conditions. The independent variable, learning conditions, had three levels: (study-study, test-test and control). The dependent variables were unit test 1, unit test 2, and unit test 3. Under the study-study (SS) condition, students received the lecture notes, and read the pre- and post-study statements before and after each of the four lessons in the unit for 15 minutes. Under the test-test (TT) condition, students received the lecture notes, took the pre- and post-test of each of the four lessons in the unit in 15 minutes, and had feedback with their answers and the correct answers at the end of each of the post-test. Under the control condition, students only received the lecture notes but did not take a test or read the study statements.

Procedure

Students received the lecture notes for the first four lessons (Control), took the pre- and post-test of the next four lessons (TT), and read the pre- and post-study statements for the last four lessons (SS). They took unit test 1 after the fourth lesson, unit test 2 after the eighth lesson, and unit test 3 after the twelfth lesson. All questions, statements or lecture notes were presented on the online course management system used by the University where the study took place. Pre-tests/study statements were open four days before the class and lasted for four days, whereas post-tests/study statements were open the day of the class and lasted for three days. Students were asked to use 15 minutes to study or take the pre-test before the class, and another 15 minutes to study or take the post-test after the class at their own time and at their own place as long as they did not seek any help in any means (people, books, notes, etc.). Lecture notes were given to students after the class.

The three unit tests, the performance scale and the helpfulness scale were also presented on the online course management system but administered by the researcher at a computer laboratory on the university campus. The same procedure was used for the three unit tests. Before each unit test, students were asked to complete the performance scale to predict how well they would perform at the unit test. Then, students took the unit test. After completing each unit test, students completed the helpfulness scale to rate how helpful the pre-/post-test, the pre-/post-study statements, or the lecture notes were in preparing them for the unit test.

Results

Unless noted otherwise, a significant level was set at .05 on all statistical tests in this study. Table 1 presents the mean unit tests scores under different learning conditions (study-study vs. test-test vs. control). Results showed a main effect of learning conditions, $F(1, 32) = 11.727$, $p < .05$, partial $\eta^2 = .268$. Further pairwise comparisons using a Bonferroni correction showed that the mean unit test score in the test-test condition (66.29%) was significantly higher than that in the study-study condition (59.47%), $p < .05$. However, the mean unit test score in the control condition (61.52%) did not differ from those in the test-test condition and study-study condition, $p > .05$.

Table 2 shows the means and standard deviations of the performance scale (1=not well at all, 10=extremely well) and helpfulness scale (1= not helpful at all, 10=extremely helpful) under different learning conditions (study-study vs. test-test vs. control). Students did not rate themselves to perform differently under different learning conditions, $F(1, 31)=.223$, $p=.64$, partial $\eta^2 = .007$. However, they did find different learning conditions helpful in preparing them for the unit tests, $F(1, 29)=6.536$, $p<.05$, partial $\eta^2 = .184$. Further pairwise comparisons using a Bonferroni correction showed that students rated the pre-/post-tests in the test-test condition (5.9) more helpful than the pre-/post-study statements in

the study-study condition (3.97). They also rated the control condition when they were given only lecture notes (5.13) more helpful than the study-study condition.

Discussion

The present study was able to generalize the *testing effect* from the laboratory to a regular college class. The mean unit test score in the test-test condition was significantly higher than that in the study-study condition. When students took the pre- and post-tests, they were able to retain more information than when they read the pre- and post-study statements. Since the time students were exposed to the pre- & post-tests and pre- & post-study statements was the same, additional exposure to the items could not explain the *testing effect*. However, the efforts students exerted when they retrieved the information to answer the unit tests further elaborate the memory traces and enhance the retention of the information (Roediger & Karpicke, 2006).

In addition, students rated the pre-/post-tests in the test-test condition and the lecture notes in the control condition more helpful in preparing them for the unit tests than the pre-/post-study statements in the study-study condition. The *testing effect* was also reflected in the helpfulness scale. Students could tell that taking a test helped them with the unit tests than studying the statements even though the study statements were actually developed from pre- and post-tests. They even found lecture notes more helpful than study statements because lecture notes covered more topics than study statements.

To enhance students' learning, college classes could incorporate frequent testing in their curriculum. The purpose of the testing is to give students opportunities to make an effort to retrieve the information from their memory. The effortful retrieval further consolidates the memory of the information. Since the purpose of the frequent testing is to provide opportunities to students to retrieve what they have learned, the testing results

should not be counted towards students' grades. In this way, students would be able to consolidate their memory of the information they have learned.

Conclusion

Testing is usually viewed as a way of assessing how much students know. However, testing can also be used to enhance students' learning. Evidence for the *testing effect* in promoting learning comes from laboratory studies and educationally related studies. With a careful design, frequent testing can be incorporated into a college class to enhance students' learning.

References

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060-1073.

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

Einstein, G. O., Mullet, H. G., Harrison, T. L. (2012). The testing effect: illustrating a fundamental concept and changing study strategies. *Teaching of Psychology, 39,* 190-193. Doi: 10.1177/0098628312450432

Faulkerson, F. E., & Martin, G. (1981). Effects of exam frequency on student performance, evaluations, of instructor, and text anxiety. *Teaching of Psychology, 8*, 90-93.

Landrum, R. E. (2007). Introductory psychology student performance: weekly quizzes followed by a cumulative final exam. *Teaching of Psychology*, *34,* 177-180.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29,* 210-212.

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.

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

Table 1

Means and Standard Deviations of the Unit Tests in Different Learning Conditions (N = 33)

		Mean	*SD*
	Study-Study (SS)	59.47%	(10.8)
	Test-Test (TT)	66.29%	(10.4)
	Control	61.52%	(11.1)

Table 2

Means and Standard Deviations of the Performance Scale (*1=not well at all, 10=extremely well) and Helpfulness Scale in Different Learning Conditions (N =* 33).

	Performed		Helpfulness	
	Mean	*SD*	Mean	*SD*
Study-Study (SS)	6.22	(1.86)	3.97	(2.66)
Test-Test (TT)	6.31	(1.89)	5.9	(2.44)
Control	6.34	(1.52)	5.13	(1.63)