CONTEMPORARY EDUCATIONAL PSYCHOLOGY 6, 217–226 (1981)

Retention of Prose Following Testing with Different Types of Tests

PHILIPPE C. DUCHASTEL
The American College

Taking a test on a passage one has just studied is known to enhance later retention of the passing contents. This study examined the effects of three types of initial test on later retention: a short-answer test, a multiple-choice test, and a full free-recall test. Questions on the first two of these tests covered only half of the passage contents. Later retention was compared for both initially tested content and untested content with that of a control group not initially tested on the passage at all. The subjects were 57 secondary school students who studied a brief history text before taking one of the initial tests. All were given retention tests 2 weeks later. The classical testing effect (enhanced retention due to initial testing) was shown to be influenced by the type of initial test used. Thus, a testing effect was evident in the case of the initial short-answer test, but not in the case of either of the other two tests. A depth-of-processing view is advanced in interpreting this finding. The testing effect was found not to generalize to untested content and in one condition (the initial multiple-choice test), retention of untested content was depressed.

A well-verified phenomenon, although one which is not currently accorded much interest, is the effect of post-testing on the retention of prose (Gates, 1917; Spitzer, 1939). This effect is described as follows: following the learning of a prose passage, a group of students who are given a post-test on the passage immediately or shortly afterward will later recall more of the passage on a retention test than will a similar group of students who are not given the initial post-test. This effect, obtained without any feedback being given to the students, can be quite remarkable, leading at times to a doubling of the retention scores for the post-tested group (e.g., Jones, 1923–1924).

Anderson and Biddle (1975) retrace the interest given to this effect over the years and attempt to provide a theoretical account of the effect in terms of a depth-of-processing hypothesis. One implication of this view, that semantic processing during the post-test would lead to greater retention later on than would verbatim processing, was tested by them in a number of studies (1975) but with only limited success. More specifically, they compared the relative effectiveness of immediate post-tests comprised of either verbatim or paraphrase questions. A testing effect was obtained with either form of questions, with the strongest effect being

Requests for reprints should be sent to Dr. Phillippe C. Duchastel, Director, Research and Evaluation, The American College, Bryn Mawr, PA 19010.

0361-476X/81/030217-10$02.00/0

obtained by a set of verbatim questions followed by paraphrase ones, i.e., a combination of both. Their failure to find paraphrase questions to be superior to verbatim ones may have resulted from their use of a multiple-choice format of testing, for it is known that a cued-recall (short-answer) retention test is the optimal format of test with which to obtain a testing effect (Anderson & Myrow, 1971).

The present study similarly examined the same depth-of-processing hypothesis, with a difference however: depth-of-processing was here defined in terms of productive memory versus identification memory, or in more familiar concepts in terms of recall versus recognition performance. Also, the retention test used in the study was a cued-recall test.

The process investigated is in general terms a process of mental review, one of reprocessing what has just been studied in attempting a test on it. The present hypothesis is that the more involved processing requirements of a cued-recall (short-answer) test over those of a recognition (multiple-choice) test will lead to comparatively greater retention later on.

Indeed, a short-answer test requires the student to retrieve the relevant information from memory and construct an answer to the question; responding to a multiple-choice question on the other hand is more strongly cued in that the correct answer is actually presented to the student as one of the alternatives. This extra cueing should reduce the degree of cognitive processing (mental review) required of the test taker and thus potentially diminish the testing effect on retention of the passage contents.

A full free-recall test contains even fewer cues to assist the student in answering the test and is therefore even more demanding in terms of cognitive processing than is a short-answer test. Without questions to act as cues however, a full free-recall test may well result in recall of only part of the passage contents. This might well result in a lesser testing effect than would be possible with a cued recall test.

Thus, the strongest testing effect was expected to result from an immediate cued-recall test; the next strongest from a recognition test, where identification of the correct alternative from a set is all that is involved; and the weakest from a full free-recall test.

Another aspect of the testing effect which was examined in this study was the effect of testing on untested material. This is similar to an examination of the effects of inserted questions on incidental material, as found in the mathemagenic literature (Rothkopf, 1976). The difference between the two research areas is that in the present case the questions all come at the end of the passage and are given in the form of an actual test.

To examine this aspect of testing, two of the tests which immediately followed learning in this study covered only half of the content of the passage. The later retention test, however, covered both previously tested and previously untested content. The question of interest was

whether the effects of testing might generalize to untested material, or whether on the contrary the retention of nontested material might be interfered with. In a study by Laporte and Voss (1975), the retention of nontested material was shown to be thus interfered with, although not greatly so (the comparison with a control group was not statistically significant). This aspect of the testing effect is an important one, for it has implications for extending the mathemagenic literature, as well as implications for the practice of testing.

METHOD

Materials

The learning passage employed in the study was developed by the author. The 1700-word passage, entitled "The Victorian Era," contained 12 topical paragraphs, each dealing with a separate topic in the period of British history covered by the reign of Queen Victoria (1837–1901). The text also contained an introduction and a conclusion; however, since these were simply meant to frame the content of the passage, only recall of the 12 topical paragraphs was examined later in the students' protocols.

The 12 topics present in the passage were the following, in their order of appearance: Prince Albert; the Corn Laws; the Great Exhibition; the Crimean War; the changing role of women; India; Livingstone; the Suez Canal; Irish Home Rule; the Queen's Jubilees; Trade-Unionism; the Boer Wars.

The passage is largely based on the account of the period given in Burke's *An Illustrated History of England* (1974). The passage was written so that it could be easily understood by the 15-year-old students taking part in the study.

Design and Subjects

The experiment involved three experimental groups, which were provided with some form of passage-related test immediately after the text was studied, and one control group, provided with a filler task. Two weeks later all students in the study received identical retention tests covering various aspects of the passage.

The first experimental group received an immediate test consisting of 12 questions requiring a brief answer to be written by the students. The following was a sample question: What nationality was Price Albert?

The second experimental group received an immediate test consisting of the same 12 questions set in a multiple-choice format with three alternatives per question. The corresponding question to the one above was:

Prince Albert was originally (a) German (b) Russian (c) Hungarian.

The third experimental group received an immediate test consisting of full free recall. The test simply requested the students to write down in their own words as many of the ideas and details that they could remember from the passage.

The filler task received by the control group students was to complete a true/false questionnaire concerned with general study habits. Their directions at the top of the questionnaire indicated that they would be given the test on the Victorian Era later on in the year and that, for the moment, they should complete the study habits questionnaire.

The names given to the groups were, in order: the Short-Answer group, the Multiple-Choice group, the Free-Recall group, and the No-Test group.

The 12 questions received by the first two experimental groups (Short-Answer and Multiple-Choice groups) related to only half of the passage contents (6 of the 12 topical

paragraphs comprising the passage). The paragraphs which were questioned on these tests were every alternate paragraph starting with the first. This experimental manipulation later enabled a comparison of the recall from questioned and unquestioned paragraphs of the text for these two experimental groups.

The subjects were 57 students in a large secondary school in Britain. They were 14 or 15 years old. The number of subjects in the experimental and control groups ranged from 10 to 17.

Retention Tests

Two tests were used as dependent measures in this study in order to reveal different aspects of retention. These were as follows in order of presentation:

Topical retention test. The students were asked to recall the 12 major topics presented in the passage and to list them in a few words each.

Cued retention test. This consisted of 36 explicit questions requiring the student to provide an idea or a detail from the passage. The following is a sample question: Who first thought of the Great Exhibition? This test was made up as follows: (i) the 12 items seen previously by the Short-Answer group and (under another form) by the Multiple-Choice group; (ii) 12 new items related to the six questioned paragraphs; and (iii) 12 items unrelated to the questioned paragraphs (and hence related to the unquestioned paragraphs). The interest for distinguishing between subsets of items which were either related or unrelated to the initial items was as follows: while both these sets of items covered previously untested material, any generalizing effect, if there were one, should be more apparent on related than on unrelated items. The items were not set out in blocks of 12 on the test, but rather followed the order of occurrence of topics in the passage.

Procedure

The study was conducted with complete classes of students during regular class time. In each class, the students were assigned to the four cells in the design by random allocation of the materials, which were contained in brown manila envelopes. Written instructions introduced the students to the task and indicated that there would be a test on the same day as well as one 2 weeks later. They were told that the test would require them to write down the main points that they could remember, and that there would be specific questions of detail as well.

The students studied the test for 15 min, after which time the immediate post-test groups proceeded with testing, while the control group proceeded with the filler task. Ten minutes time was allowed for this testing phase. This amount of time may have been too little for some of the students in the Free Recall group but it was ample time for the students in the three other groups. This issue will be taken up later in the discussion section of the paper.

The retention tests were administered 2 weeks later to all students. These tests were corrected in a blind manner, whereby the person correcting the tests (a research assistant) was not aware of the group identity of the pupils as she examined their answers.

The message had been collected after the students had studied it and was therefore not available to the students during the interval between initial testing and retention testing. There is no evidence concerning the degree to which the students may have talked about the contents of the passage among themselves in the two-week interval, but presumably this random factor would be spread out across all groups.

RESULTS

Two of the tests administered during the first session as part of the treatments were scored to obtain some indication of how well the students

were doing and to verify that they had taken the task seriously. The two tests scored were the short-answer test and the multiple-choice test. The means on these two tests were 8.2 (SD-2.0) and 9.6 (SD-2.3), respectively, both out of a maximum of 12. Thus the students were scoring satisfactorily and as expected, the multiple-choice scores were somewhat higher than the short-answer scores. The free-recall test was not scored for lack of an objective scoring grid and because scoring of the other two tests sufficed in establishing the point just made.

Testing Effect

The first retention test administered to all students was the topical retention test in which the students were requested to list the major topics they encountered in the passage during the first session. The means and standard deviation scores from this test are presented in the first part of Table 1. Also presented in the table are the means and standard deviations of the scores obtained in each group on the cued retention test. This test contained 36 questions, each requiring a short answer.

The scores on these two tests are rather low (under 50% in all cases) but that is not too surprising for a retention test. An analysis of variance was performed on each of these two tests with the following results: only the analysis of the cued retention scores led to a significant result, $F(3.53) = 3.4$, $p < .05$.

Multiple comparisons using the Duncan procedure revealed that the Short-Answer group differed significantly from the control group on the cued-retention test. The means reveal that the Short-Answer group scored 30% higher than the Control group on this retention test. The Short-Answer group was also found to be significantly superior to the two other groups on this test.

Mathemagenic Effects

As indicated earlier, one of the purposes of this study was to explore the effect of testing on untested material. To this end, the topical retention

TABLE 1
MEANS AND STANDARD DEVIATIONS ON THE TWO RETENTION TESTS, BY GROUP

		Topical retention (Max:12)		Cued retention (Max:36)	
	N	*M*	SD	*M*	SD
Short-answer group	17	4.4	2.5	14.5	5.6
Multiple-choice group	14	2.7	2.0	10.1	3.6
Free-recall group	10	4.3	2.0	9.1	3.6
No-test (control) group	16	3.3	2.4	11.1	5.7

scores were partitioned into two subscores: the first of these related to the content in the passage which was tested initially and which was being retested here; the second subscore related to passage content not tested during the first session.

Similarly, the cued retention scores were partitioned into three subscores: the first was based on repeated questions (seen previously on the initial test by the Short-Answer and Multiple-Choice groups); the second subscore was based on related questions (pertaining to the same six passage paragraphs as the repeated questions); and the third subscore was based on unrelated questions (pertaining to the six paragraphs on which the Short-Answer and Multiple-Choice groups had not been previously tested).

All of these subscores are presented in Table 2, in terms of means and standard deviations for each of the four groups. Analyses of variance were performed on each of these subscores. The analyses of the two topical retention subtests did not lead to significant differences between the groups. The analyses of the cued retention subtests, on the other hand, did lead to significant results: both the Repeated Questions analysis and the Related Questions analysis were significant, $F(3.53) = 5.9, p < .01$ and $F(3.53) = 2.8, p < .05$, respectively. The analysis of the Unrelated Questions scores was not significant although close ($p = .10$).

Multiple comparisons based on the Duncan procedure between each of the experimental groups and the control group revealed some reliable differences in group scores. Thus the Short-Answer group scored significantly more (65% higher) on the Repeated Questions than did the No-Test group. Also the Multiple-Choice group performed significantly worse (55% less) than the Control group on the Unrelated Questions. No other comparisons with the control group led to a significant difference. Thus on the cued retention test, the Short-Answer treatment enhanced the learning of tested material whereas the Multiple-Choice treatment depressed the learning of unrelated material (the six paragraphs on which no questions had been posed during the initial test).

Multiple comparisons also revealed that the Free-Recall group performed significantly worse than the other two experimental groups on the Repeated Questions and that the Short-Answer group performed significantly better than its two counterpart experimental groups on Related Questions.

DISCUSSION

The testing effect hypothesis elaborated in this study predicted that different types of tests would differentially affect retention, with the largest effect being obtained by the group initially tested with a short-answer test. This prediction was supported by the data in the case of the cued

TABLE 2

MEANS AND STANDARD DEVIATIONS OF THE SUBSCORES DERIVED FROM THE TWO RETENTION TESTS, BY GROUP

	Topical recall				Cued recall					
	Tested topics (Max:6)		Untested topics (Max:6)		Repeated questions (Max:12)		Related questions (Max:12)		Unrelated questions (Max:12)	
	M	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Short-answer group	2.9	1.5	1.4	1.4	6.7	2.5	5.6	1.9	2.2	2.2
Multiple-choice group	2.0	1.3	0.7	1.3	5.2	2.0	3.9	1.7	1.0	1.0
Free-recall group	2.7	.9	1.6	1.7	3.1	1.7	3.9	1.9	2.1	.8
No-test (control) group	2.2	1.6	1.1	1.2	4.0	2.9	4.7	2.1	2.4	1.6

retention test but not in the case of the topical retention test. This latter test is admittedly the less sensitive of the two and this, combined with the low number of subjects in some groups, may help to explain why no testing effect was obtained in this case.

It is on the cued retention test scores that the testing effect comes out most clearly. It also points to differential effects. On this retention test, a mild testing effect (30% increase) was revealed only in the case of the prior short-answer test and not with either the prior multiple-choice test or the prior free-recall test. It is also interesting to note that initial testing with a short-answer test was superior to such testing with either a multiple choice or a free-recall test. Thus an important initial finding of this study is that the testing effect can be strongly influenced by the type of test which follows learning.

The results of Anderson and Biddle (1975) also support this contention, but contrary to their study, the present findings can be interpreted as supporting a depth-of-processing explanation. Indeed, as predicted, the processing requirements of a short-answer test proved more effective for retention than those of a multiple-choice test. The presumed reason for this is that the retrieval from memory of appropriate answers in response to open questions (short-answer test) involves a deeper memory search than does the identification of appropriate alternatives in answering multiple-choice questions. In other words, in the first case, the processing of what one has just learnt must be carried out to a deeper level (in terms of memory search) than in the second case. In this respect, however, a better explanatory term than depth-of-processing might be depth-of-*re*processing, for in a testing situation a student is not processing new information but rather making use of old information. An even more preferred term would be the classical one of consolidation (Jones, 1923–1924).

A competing explanation for the obtained results might be advanced in the form of a practice effect, since the cued retention test was similar in format to the initial test received by the Short-Answer group but not to the initial tests received by the other groups. Thus, the similarity of the initial and retention tests in the case of the Short-Answer group might have contributed to the outcomes. Practice itself, however, does not adequately explain the differential outcomes between the groups which were obtained on the various subtests comprising the cued retention test. Some form of practice may, however, be involved to a certain degree in the effects obtained and the issue remains worthy of further investigation (current research by the author is aimed at examining this issue).

The relative inefficiency of the full free-recall test in enhancing retention in this study was likely due to two factors: (a) the greater difficulty in responding to the test, since no cues in the form of specific questions were

provided as in the case of the other two types of test; and (b) the less than optimal time limit which was imposed (only 10 min). This time limit was imposed on all groups for purely practical reasons related to the conduct of the study and since it most likely disadvantaged the group in question, little can be reliably concluded at this time about the relative value of a free-recall test in enhancing later retention.

A testing effect was not obtained in all cases where it was expected in this study, nor was it particularly strong when obtained with the short-answer test. One reason for this must be the fact that the total retention scores discussed up till now were in fact constituted of distinct subscores each covering very different material. Some parts of the material had been tested before while other parts had not been tested. For indeed, an additional intent of the study had been to explore the effects of testing on untested material.

Each of the two tests had been partitioned into different types of subtests. Thus the topical retention test consisted of six elements related to previously tested content and six elements related to untested content. This test, however, was not sensitive enough to create differences between the groups on either subtest.

The cued retention test had been partitioned into three subtests: repeated questions, related questions (covering content from the same paragraphs to which were geared the repeated questions), and unrelated questions (covering content from the other paragraphs). It should be realized that these distinctions apply strictly only in the case of the Short-Answer and Multiple-Choice groups, since the other two groups received as part of their treatments either a filler task or a full free-recall test covering the whole passage. Our interest therefore lies with the performance of the first two experimental groups mentioned above in relation to that of the control group.

For the control group, the three subscores have little differential meaning and should ideally have been approximately equal. The large differences between them simply reflect differential item set difficulties, which is of little consequence per se since all comparisons are made within each set across groups.

The repeated questions analysis revealed a strong testing effect in the case of the Short-Answer group. The related questions analysis indicated no effect in either group and the unrelated questions analysis revealed a depression of scores for the Multiple-Choice group.

This pattern of results is not unlike a pattern found in the mathemagenic research literature where relevant learning (equivalent here to the repeated questions) is generally enhanced and incidental learning (equivalent here especially to the unrelated questions) is sometimes depressed by questions inserted in text. Thus testing after the learning of a brief text

would seem not to have a generalizing effect but rather to have a focusing one, whereby the retention of tested material is enhanced although possibly at the expense of untested material.

This remains but a tentative generalization, however, as the interpretation of the data obtained in the present study is not unequivocal. The fact that the testing effect was evident on repeated questions for the Short-Answer group makes strong sense in the light of our earlier discussion of the relative effectiveness of different types of test. It is less clear, however, why depressed retention of untested material was particularly obtained in the case of the Multiple-Choice group and not in the case of the Short-Answer group. A plausible explanation remains wanting at this time.

In summary, this study has demonstrated initially that different types of tests which follow the learning of a prose passage can differentially affect the degree of the testing effect to be derived from such a practice of testing. This differential effect can usefully be interpreted in terms of the depth-of-processing required by the different types of tests.

This study also provides evidence which supports the notion that the effect of testing is not likely to generalize to untested content and, more tentatively, that under certain conditions testing may even depress the learning of content which was not tested. This problem is an important one in terms of pedagogical practice and merits further study.

REFERENCES

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

ANDERSON, R. C., & MYROW, D. L. Retroactive inhibition of meaningful discourse. *Journal of Educational Psychology Monograph*, 1971, **62**, 81–94.

BURKE, J. *An illustrated history of England*. London: Collins, 1974.

GATES, A. I. Recitation as a factor in memorizing. *Archives of Psychology*, 1917, **6**, 1–104.

JONES, H. E. The effects of examination on permanence of learning. *Archives of Psychology*, 1923–1924, **10**, 1–70.

LAPORTE, R. E., & VOSS, J. F. Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, 1975, **67**, 259–266.

ROTHKOPF, E. Writing to teach and reading to learn: A perspective on the psychology of written instruction. *Seventy-Fifth Yearbook of the National Society for the Study of Education, Part 1*, 1976, 91–129.

SEHULSTER, J. R., & CROUSE, J. H. Storage and retrieval of prose material. *Psychological Reports*, 1972, **30**, 435–439.

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