Journal of Educational Psychology
1982, Vol. 74, No. 1, 18-22

0022-0663/82/7401-0018$00.75

Testing Versus Review: Effects on Retention

Ronald J. Nungester and Philippe C. Duchastel
The American College

Taking a test on content that has just been studied is known to enhance later retention of the material studied, but is testing more profitable than the same amount of time spent in review? High school students studied a brief history text, then either took a test on the passage, spent equivalent time reviewing the passage, or went on to an unrelated task. A retention test given 2 weeks later indicated that the test condition resulted in better retention than either the review or the control conditions. The effect was further shown to be content specific (in contrast to effects typically produced by questions inserted in text) and independent of item format. These results favor a greater use of testing in instruction.

Administering quizzes to students in class is generally considered to fulfill two functions: to motivate students to study and to determine how well they have mastered the material that was taught. A third function, more directly related to the learning process, goes largely unrecognized: to help the student consolidate in memory what was learned. It is this third function of testing with which the present research is concerned.

This consolidation function of testing was demonstrated relatively early in instructional psychology (Jones, 1923–1924) and replicated on numerous occasions (e.g., Laporte & Voss, 1975). This consolidation effect is described as follows: taking a test immediately after learning will lead to better retention of the material at a later date, as evidenced on a delayed retention test, even when no corrective feedback is provided and when no further study of the material has taken place.

Recent research (Duchastel, 1981; Nungester & Duchastel, Note 1) has examined how this consolidation effect (known simply as a testing effect on retention) was influenced by the type of test employed. Two test formats were considered: short-answer tests and multiple-choice tests. This research initially indicated an advantage for short-answer tests but later demonstrated that multiple-choice tests can be just as potent for enhancing retention. Thus, at the moment, there is no strong basis for concluding that one type of test has the advantage over the other for consolidating learning. Our previous research (Nungester & Duchastel, Note 1) has also shown that the consolidation effect is independent of the simple test practice effect derived from repeated testing with the same type of test. That is, a testing effect can also be demonstrated on a retention test cast in a different format (e.g., on a multiple-choice retention test when the initial test was a short-answer one).

Of practical concern to teachers is the question of whether the time devoted to testing might be spent as profitably by allowing students to study the material more. Is spending some portion of a teaching session in testing really more valuable than spending that same time in further study? The present experiment principally addressed this applied question.

From a learning process point of view, this experiment examined the possibility that observed testing effects are not due to testing itself but result from the fact that experimental groups spend more total time on a topic (learning time and testing time) than do the control groups typically employed (who spend the same amount of time on the learning task, but then go on to some other

We wish to acknowledge the assistance of the Haverford Township School District, Delaware County, Pennsylvania, in the conduct of this study, especially the teachers who assisted directly: Mr. Bush, Mrs. McGarvey, Miss Harrison, Mr. Long, and their principal, Mr. Drukin.

Requests for reprints should be sent to Ronald J. Nungester, The American College, 270 Bryn Mawr Ave., Bryn Mawr, Pennsylvania 19010.

task used as a filler task, such as completing a study habits inventory or the like). This argument is called the total-time hypothesis and has been invoked in the connex area of research on adjunct aids as a competing explanation for the results obtained in many experimental situations such as these (Faw & Waller, 1976). The total-time hypothesis thus vies with the consolidation hypothesis as an explanation for the effects of testing on later retention. The design of the present study allowed these two hypotheses to confront one another and thus offered a serious test of the consolidation hypothesis.

Three groups of students studied a brief history text, after which the first group was tested on the passage, the second group was allowed further study of the text for an equivalent amount of time, and the third group was directed to an unrelated (filler) task. A retention test on the passage was administered to all students 2 weeks later. It was expected that performance on this retention test would be strongest for the group initially tested following learning, next strongest for the group allowed further study, and weakest for the filler task group.

A further refinement in the design (described in the next section) permitted a replication of our previous findings with respect to test format, as well as an examination of how the testing of some content might affect the later retention of other, initially untested, content. The primary aim of the experiment, however, was to contrast testing with further studying, as indicated above.

Method

Subjects

The students participating in the experiment were 97 senior students from a middle-class suburban high school. They participated in the experiment as part of their regular school program. The students were randomly assigned to the three conditions in the study.

Materials

The learning passage employed in this study was the same one that was employed in the previous two studies by the authors. It consisted of a 1,700-word passage entitled "The Victorian Era," which contained 12 topical paragraphs describing events in British history (1837–1901 period). The passage had been adapted from other sources by one of the present authors so that it could be easily understood by high school students. The passage is more fully described by Duchastel (1981).

Design and Procedure

The experiment involved two experimental groups and one control group. The first group, the test group, studied the passage for 15 minutes, then took an immediate test on its contents (initial test). No feedback was provided. The second group studied the passage for 15 minutes, then spent additional time reviewing the passage. This group was labeled the review group. The control group studied the passage for 15 minutes, then completed a learning process questionnaire that served as a filler task. This filler task simply served to occupy the students in this group while the other students were either completing the initial test or reviewing the passage. The time allocated for either treatment (test or review) or for the filler task was 5 minutes.

Two weeks later, all students were administered a retention test on the contents of the materials.

The history passage used in the experiment was collected after the students had initially studied it and was therefore not available to the students during the interval between the two experimental sessions. Their teachers were furthermore asked not to discuss this part of history with them until after the retention test. A questionnaire administered to the students at the conclusion of the experiment inquired about any discussion of the text with friends during the 2-week interval.

The teachers were aware of an eventual retention test, but the students themselves were not told of such a test. To provide some apparent conclusion to the experiment at the end of the first session, the students were administered a brief elaborative processing inventory, developed by Schmeck, Ribich, and Ramanaiah (1977). Bringing closure to the initial session in this way was especially important for the review and control groups, since they were not tested on the content of the passage during this session.

Tests

The initial test, which was administered to the students in the test group only, contained 12 questions that selectively sampled the contents of the passage. Every odd-numbered question was in a multiple-choice format (e.g., "What nationality was Prince Albert? a) German; b) Russian; c) Hungarian."); and every even-numbered question was in a short-answer format (e.g., "In which part of the world was the Crimean War? _____"). Each of the 12 questions corresponded to one of the 12 topics in the passage.

The retention test, which was administered 2 weeks later to all students in the study, contained 24 questions (two questions per topic). For the test group, half of these questions were old questions that required recall or recognition of the same information as requested on the initial test. The other half of the questions were new questions for this group. For the review and con-

trol groups, all questions on the retention test were in fact new questions, since neither of these groups were tested in the initial session.

The questions the test group had seen on the initial test (old questions) were transformed into the alternate question format on the retention test. Thus, multiple-choice questions on the initial test became short-answer questions on the retention test, and vice versa. To the illustrative questions presented above corresponded the following questions: "What nationality was Prince Albert? _____" and "In which part of the world was the Crimean War? a) the Near East; b) North-Africa; c) India." Reversal of item format in this way enabled us to replicate some of the conditions found in our previous study (Nungester & Duchastel, Note 1).

As can be seen from the illustrative questions above, all questions were at the information level of knowledge.

Results

The retention test scores are presented in Table 1. With respect to the total test scores, the pattern of results indicated that the test group performed best of all, followed by the review group and then the control group. An analysis of variance performed on these scores revealed a significant difference, $F(2, 94) = 4.0$, $p < .05$, but further planned contrasts between each pair of groups revealed that only the difference between the test group and the control group was statistically significant, $p < .05$. The sample difference between the test group and the review group was not significant. These results have implications for the total-time hypothesis and are discussed in the next section.

In the first part of Table 1, the total test scores are partitioned according to whether the questions represent new or old items in terms of the prior experience of the test group. For the other two groups, all questions were new ones and the partition only serves to provide baselines with which to compare the two subsets of items identified in the case of the test group.

Table 1
Means and Standard Deviations of Scores on the Retention Test (Subsets of Items and Total Test)

	Subset A		Subset B		Total	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test[a] ($n = 31$)	7.7	2.3	4.7	2.7	12.4	4.5
Review ($n = 34$)	5.7	1.8	5.3	2.0	11.0	3.1
Control ($n = 32$)	5.0	2.1	4.6	2.5	9.7	3.9

Note. The test contained 24 items.
[a] For this group only, Subset A represents items repeated from the initial test; Subset B represents new items.

Table 2
Means and Standard Deviations of the Retention Test Scores Partitioned According to Question Format (Total Test and Old Items Only)

	Total test				Old items only			
	MC		SA		MC		SA	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test	7.1	2.3	5.3	2.6	4.3	1.3	3.5	1.3
Review	6.7	1.6	4.3	2.0	3.4	.9	2.4	1.2
Control	6.0	2.2	3.7	2.5	3.2	1.3	1.8	1.3

Note. MC = multiple-choice questions; SA = short-answer questions.

Analyses of variance were performed on both subsets of items and revealed that a significant difference existed in the case of old items but not in the case of new items, $F(2, 94) = 14.6$, $p < .001$, and $F(2, 94) < 1$, respectively. Planned contrasts within the subset of old items revealed that the test group differed significantly from both the review and control groups ($p < .001$, in each case). Thus, the benefits of testing were limited to old items and did not extend to new ones.

Another way of partitioning the total test scores is in terms of the format of the questions: multiple-choice or short-answer. The partitioned scores are presented in Table 2. Analyses of variance on each set of questions for the total test revealed that a significant difference existed only in the case of the short-answer questions, $F(2, 94) = 3.9$, $p < .05$. However, when only old items were considered (these being the only items that revealed a testing effect in the previous analysis), the partition revealed that a significant difference existed in both the case of short-answer questions and of multiple-choice questions, Fs(2, 94) = 13.7 and 7.3, $p < .001$ and $p < .05$, respectively. Thus, testing effects were not limited by item format.

Finally, the brief questionnaire administered to the students at the end of the study indicated that a number of students thought about the text contents or discussed them among themselves in the intersession interval. The proportion of students who did so ranged from 60% (test and control groups) to 70% (review group). To examine how this intersession activity might have influenced the retention results, the data were re-analyzed with the students partitioned into those who did discuss the text contents and those who did not. The results of these analyses did not differ from those reported above. Furthermore, the correlation between the initial test and the retention test calculated in the case of the test group was .78. Thus, although intersession activity may have slightly increased overall retention performance, it did not do so differentially between the groups.

Discussion

The principal aim of this study was to examine a practical issue concerning the testing effect: Should students spend some portion of learning time on testing or simply devote that time to study or continued review? The results of the study indicate that testing is indeed more profitable for retention.

Although review itself is profitable, as indicated by a sample increase of 10% in retention over control group performance (total test scores), testing is even more profitable (resulting in a sample increase of 25% over control group performance). Testing thus appears to have the advantage.

This decision-oriented conclusion may seem to be at odds with the fact that the contrast between testing and review on total test performance was not statistically significant. This, however, was true only for total test performance. When subsets of the test questions were examined in terms of the old versus new items for the test group, the results were different: The groups do not differ on new items, but the test group is superior to both other groups on old items. It is this particular result that leads us to conclude that testing has a definite advantage over review, as explained below.

The design of the study called for initial testing with only half of the items that constituted the retention test. It is on these items that students in the test group showed an advantage over review students on the retention test. Had the initial test comprised all of the items on the retention test, it is most likely that this advantage would have been evident in total test performance. We therefore feel that it is warranted at this time, given the difference between groups on old items, to conclude that testing is indeed more advantageous for retention than is review.

This same interpretation also extends to the more theoretical issue concerning the total-time hypothesis. The previous research on testing had shown that testing can enhance retention, but no account had been taken of the additional time required for testing. The present experiment demonstrated that testing remains beneficial even when such testing replaces actual study (review) time. The total-time hypothesis therefore does not limit the validity of the testing effect, nor does it limit the applicability of the testing principle to actual practices in school settings.

It remains possible of course that more difficult or complex texts requiring greater comprehension skills would profit more from additional study than did the factually oriented text used in this study. Until the generalizability of the present results are further examined, conclusions should be restricted to the testing of factual materials.

Whereas the focus of this experiment was the practical issue discussed above, the design was additionally motivated by a desire to partially replicate our previous findings with respect to the retention of content initially untested (Duchastel, 1981) and with respect to test format (Nungester & Duchastel, Note 1).

Our design decision to employ a retention test that comprised both items seen before by the test group (old items) and items not previously seen (new items) was aimed at determining whether testing has a specific or a general effect in terms of consolidation. That is, are only contents covered by the initial test in fact consolidated, or do other contents in the passage also share in this

process of consolidation, even though they are not represented on the initial test?

In the mathemagenics literature, both specific and general processes have been demonstrated to result from inserted post questions (McGaw & Grotelueschen, 1972; Rickards, 1979). In the literature dealing with the testing effect however, only a specific process has been demonstrated: Both Laporte and Voss (1975) and Duchastel (1981) have found that the testing of specific content enhances retention of that content, but does not enhance the retention of other, initially untested, content. The present findings further support this conclusion: Old items on the retention test were better answered by the test group when compared with the other two groups, but not new items. Consolidation would thus appear to be limited to the contents of the passage that are tested. This conclusion points to a major area of divergence between the mathemagenics literature and the testing effect literature.

A second design decision in this study was to employ both short-answer and multiple-choice questions on the initial test and to reverse the test format of these items on the retention test. This arrangement does not permit a true experimental test of the item format issue (for lack of a control condition for which item format would not be reversed), but it does permit a partial replication of our previous finding that the testing effect is not fully confounded with a test practice effect (Nungester & Duchastel, Note 1). This replication was positive: A testing effect was obtained with our items even though test format was reversed. We are therefore more confident in our previous conclusion that the testing effect does indeed involve consolidation and is not merely an artifact of a repeated test format.

The replication also confirmed our earlier result that testing effects are not greatly influenced by initial item format; indeed, testing effects are as readily obtained with multiple-choice items as with short-answer ones.

Practical Considerations

The previous research on testing had established that testing can be a potent way of enhancing retention. The present study demonstrated that testing is superior to review for that purpose; thus, although testing takes time away from study, that time is well spent.

It should be noted that the present study involved unguided review in the form of additional study time. Directed review activities that structure the review, whether encouraged by written instructions following the text or led by a teacher, would possibly attenuate the advantage of testing. Directed review might in fact serve in this respect the consolidation function offered by testing.

As indicated in our introduction, educators are apt to value testing (in the form of quizzes) for motivational and diagnostic purposes. The research on the testing effect adds a further dimension to the use of quizzes. As such, it should encourage educators to make greater use of testing in instruction.

Reference Note

1. Nungester, R. J., & Duchastel, P. C. *Testing effects measured with alternate test forms.* Paper presented at the meeting of the American Educational Research Association, Los Angeles, 1981.

References

Duchastel, P. Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, 1981, *6*, 217–226.

Faw, H. W., & Waller, T. G. Mathemagenic behaviors and efficiency in learning from prose. *Review of Educational Research*, 1976, *46*, 691–720.

Jones, H. E. The effects of examination on the performance of learning. *Archives of Psychology*, 1923–1924, *10*, 1–70.

Laporte, R., & Voss, J. Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, 1975, *67*, 259–266.

McGaw, B., & Grotelueschen, A. Direction of the effect of questions in prose material. *Journal of Educational Psychology*, 1972, *63*, 580–588.

Rickards, J. Adjunct postquestions in text: A critical review of methods and processes. *Review of Educational Research*, 1979, *49*, 181–196.

Schmeck, R., Ribich, F., & Ramanaiah, N. Development of a self-report inventory for assessing individual differences in learning processes. *Applied Psychological Measurement*, 1977, *1*, 413–431.

Received January 27, 1981 ■