Effects of Anticipation of Tests on Delayed Retention Learning

W. J. Haynie, III

The benefits of tests as aids to learning, beyond their primary evaluation function, have been studied in a variety of settings. This study sought to isolate the effects of anticipation of a test (and the assumed improvement in study and preparation commensurate with such anticipation) from the learning gains resulting from the act of taking the test. The investigation involved instruction via self-paced texts, initial testing of learning, and delayed testing three weeks later. The delayed tests provided the experimental data for the study. The investigation also included a survey to determine perceptions of students concerning classroom tests.

Background

Most of the research on testing which has been reported in recent years has concerned standardized tests, but much of the evaluation done in schools is done with teacher-made tests (Haynie, 1983, 1990a; Herman & Dorr-Bremme, 1982; Mehrens, 1987; Mehrens & Lehmann, 1987; Newman & Stallings, 1982; Stiggins, Conklin, and Bridgeford, 1986). Research is needed on the effects of teacher-made tests and other issues surrounding them such as frequency of use, quality, benefits for student learning, optimal types to employ, and usefulness in evaluation. The available findings on the quality of teacher-made tests cast some doubt on the ability of teachers to perform evaluation effectively (Carter, 1984, Fleming & Chambers, 1983; Gullickson & Ellwein, 1985; Haynie, 1992, 1995b; Hoepfl, 1994; Stiggins & Bridgeford, 1985). Despite the recognized faults, Mehrens and Lehmann (1987) point out the importance of teacher-made tests in the classroom and their ability to be tailored to specific instructional objectives. Evaluation by teacher-made tests in schools is an important and needed part of the educational system and a crucial area for research (Ellsworth, Dunnell, & Duell, 1990; Haynie, 1990a, 1992; Mehrens & Lehmann, 1987; Nitko, 1989).

The effectiveness of test taking as an aid to retention has been studied in several settings and in association with several related variables. In all of these studies, test taking has been shown to aid retention of learned material (Haynie

---

W. J. (James) Haynie, III is an Associate Professor in the Department of Math, Science, and Technology Education, University of North Carolina, Raleigh, North Carolina.

1990a, 1990b, 1991, 1994, 1995a; Nungester & Duchastel 1982). Reviewers of some earlier works which used the general protocol of this study to examine the benefits of various types of tests and methods of testing/reviewing as aids to retention criticized the works by pointing out that experimental groups in many of the studies expected to be tested whereas the control groups did not. The logical argument was that students in the experimental groups paid more attention to the study of the material and thus, it was difficult to separate the gains made while studying more diligently from those claimed by the investigators to result from the act of taking the test (testing effect). Only one of those studies demonstrated a clear separation of these two factors (Haynie, 1990a), and it was conducted in a secondary school setting with videotaped materials as the teaching-learning method. Another criticism of the protocol has been that students did not expect the test scores to be counted in determination of their course grades, so they may not have taken the entire unit of instruction seriously. Lastly, in most of the earlier studies, no attempt was made to insure equal ability of the groups other than randomization of treatment assignment. This investigation examined some of the same questions as earlier studies with careful attention to address these criticisms.

Purpose and Definition of Terms

The purpose of this study was to investigate the effects of anticipation of an upcoming test and the act of taking a test as aids to retention learning. "Retention learning" as used here refers to learning which lasts beyond the initial testing and it is assessed with tests administered 2 or more weeks after the information has been taught and tested (Haynie, 1990a; Nungester & Duchastel, 1982). A delay period of three weeks was used in this study. "Initial testing" refers to the commonly employed evaluation by testing which occurs at the time of instruction or immediately thereafter. "Delayed retention tests" are research instruments which are administered 2 or more weeks after instruction and initial testing to measure retained knowledge. (Duchastel, 1981; Haynie, 1990a, 1990b, 1991, 1994, 1995a; Nungester & Duchastel, 1982). The delayed retention test results were the only data analyzed in the experimental portion of this investigation. Additionally, one group was asked to respond to a questionnaire concerning classroom testing. The responses were analyzed and are reported in this article.

The research questions posed and addressed by this study were:

1. If delayed retention learning is the objective of instruction, does initial testing of the information aid retention learning?
2. If delayed retention learning is the objective of instruction, does the anticipation of an upcoming test on the information aid retention learning?
3. Do students study with greater effort when they expect a test than when they do not expect a test?

Methodology

Population and Sample

Undergraduate students in 6 intact technology education classes were provided a booklet on new "high-tech" materials developed for space exploration. There were 110 students divided into three groups: (a) Test Announced, Test Given (Group A, *n*=37), (b) Test Announced, No Test Given (Group B, *n*=35), and (c) No Test Announced, No Test Given (Control, Group C, *n*=38). All groups were from the Technology Education metals technology (TED 122) classes at North Carolina State University. Students were freshmen and sophomores in Technology Education, Design, or in various engineering curricula. Students majoring in Aerospace Engineering were deleted from the final sample because much of the material was novel to other students but had previously been studied by these students.

Group assignment to instructor was not randomized due to scheduling restraints, however, all sections were taught by either the researcher or his graduate assistant—each teaching some control and some experimental sections. The course instructor gave no instruction or review to any groups and provided the directions for participation via a scripted standard statement. Two sections were in each group. Random assignment of groups to treatments, deletion of students majoring in Aerospace Engineering, and absences on testing dates resulted in final group sizes which were unequal. To establish equality of ability prior to conduct of the study, the means of the first subtest taken in the course were compared. This subtest on precision measurement, metallurgy, and sheet metal processes comprised the Metals Pretest.

Design

At the beginning of the course it was announced that students would be asked to participate in an experimental study and that they would be learning subject matter reflected in the newly revised course outline while doing so. The two experimental groups were both told that the test they would be given on this new material would be counted equally with the other tests in determining course grades. The control group, however, was told that formal tests had not been prepared on the added material, so this portion of the course would not be considered when determining course grades except to insure that they made a "good, honest attempt". All other instructional units in the course were learned by students working in self-paced groups and taking subtests on the units as they studied them. The subtests were administered on three examination dates. The experimental study did not begin until after the first of the three examination dates to insure that students could see (and believe) that none of the eight subtests reflected the newly added subject matter. Students' scores on the first subtest (Metals Pretest) were compared to insure that the groups were of equal ability.

During the class period following the first examination date, the subtests which had been taken were reviewed and instructions for participation in the experimental study were given. All students were given copies of a 34 page study packet prepared by the researcher. The packet was titled "High

Technology Materials" and it discussed composite materials, heat shielding materials, and non-traditional metals developed for the space exploration program and illustrated their uses in consumer products. The packet was in booklet form. It included the following resources typically found in textbooks: (a) A table of contents, (b) text (written by the researcher), (c) halftone photographs, (d) quotations from other sources, (e) diagrams and graphs, (f) numbered pages, (g) excerpts from other sources, and (h) an index with 119 entries correctly keyed to the page numbers inside. Approximately one-third of the information in the text booklet was actually reflected in the tests. The remainder of the material appeared to be equally relevant but served as a complex distracting field to prevent mere memorization of facts. Students were instructed to use the booklet as if it were a textbook and study as they normally would any class assignment.

Group A and Group B were both told to study the packet and they would be tested on the material in-class two weeks later. Both groups were requested to return the packets on the test date also. Students were told that the results would be used along with other subtest scores in determining their course grades. On the announced test date, Group A was actually administered the initial posttest, but Group B was asked to complete a questionnaire instead. Group B was then told that the test was not ready and so their highest subtest scores would be counted double in determining their grade.

In order to obtain a control group, two sections of students in the same course were given similar initial instructions, but they were not told they were in an experiment. They were merely told that the material was newly added to the course and no subtests had been prepared yet—so they were simply lucky and would be expected to study the material as if they would be tested, however, they would not actually be tested. These students comprised Group C (control).

Three weeks later, all groups were asked to take an unannounced delayed retention test on the same material. They were told at this time that the true objective of the experimental study was to see which type of test (or no test) promoted delayed retention learning best, and that their earlier tests, if any, were not a part of the study data in any way. They were asked to do their best and told that it did not affect their grades. Participation was voluntary, but all students did cooperate.

The same room was used for all groups during instructional and testing periods and while directions were given. This helped to control extraneous variables due to environment. The same two teachers provided all directions (from prepared scripts) and neither administered any instruction in addition to the texts. Students were asked not to discuss the study or the text materials in any way. All class sections met for 2 hours on a Monday-Wednesday-Friday schedule. Half of the students in each group were in 8:00 a.m. to 10:00 a.m. sections and the others were in 10:00 a.m. to 12:00 noon sections, so neither time of day nor day of the week should act as confounding variables. Normal precautions were taken to assure a good learning and testing environment.

Instrumentation

The initial test was a 20 item multiple-choice test. The items had five response alternatives. The test operated primarily at the first three levels of the cognitive domain: Knowledge, comprehension, and application.

The delayed retention test was a 30 item multiple-choice test. Twenty of the items in the retention test were alternate forms of the same items used on the initial test. These served as a subtest of previously tested information. The remaining ten items were similar in nature and difficulty to the others, but they had not appeared on the initial test. These were interspersed throughout the test and they served as a subtest of new information.

The delayed retention test was developed and used in a previous study (Haynie, 1990a). It had been refined from an initial bank of 76 paired items and examined carefully for content validity. Cronbach's Coefficient Alpha procedure was used to establish a reliability of .74 for the delayed retention test. Item analysis detected no weak items in the delayed retention test.

Data Collection

Students were given initial instructions concerning the learning booklets and directed when to return the booklets and take the test. The test (Group A) or questionnaire (Group B) was administered on the same day that the booklets were collected. The unannounced delayed retention test was administered three weeks later. Data were collected on mark-sense forms from National Computer Systems, Inc.

Data Analysis

The data were analyzed with SAS (Statistical Analysis System) software from the SAS Institute, Inc. The answer forms were electronically scanned and data stored on floppy disk. The General Linear Models (GLM) procedure of SAS was chosen for omnibus testing rather than analysis of variance (ANOVA) because it is less affected by unequal group sizes. A simple one-way GLM analysis was chosen because the only experimental data consisted of the Delayed Retention Test means of the three groups. This procedure was first applied to the first regular subtest given in the course (Metals Pretest) to determine if groups had equal entering ability. The GLM procedure was then used again with the Delayed Retention Test means. Follow-up comparisons were conducted via Least Significant Difference *t*-test (LSD) as implemented in SAS. Alpha was set at the $p<.05$ level for all tests of significance. Tabulations of frequency and percentage were the only analysis of the survey data.

Findings

The means, standard deviations, and final sizes of the three groups on the Metals Pretest and the Delayed Retention Test are presented in Table 1. The overall difficulty of the Delayed Retention Test can be estimated by examining the grand mean and the range of scores. The grand mean of all participants was 15.85 with a range of 6 to 27 on the 30 item test. No student scored 100% and the grand mean was close to 50%, so the test was relatively difficult. The grand mean, however, was not used in any other analysis of the data.

The GLM procedure was used to compare the 3 groups on the Metals Pretest to determine if they were equal in ability prior to participating in the experimental portion of the study. The means appear in Table 1. A finding of F (2, 107) = 0.29, p = .748 indicated that the groups were equal in their entering ability (Table 2).

Table 1
Means, standard deviations, and sample sizes

	Metals Pretest		Delayed Retention Test	
Treatment	Mean	*SD*	Mean	*SD*
Group A Test Announced/Given *n*=37	22.5	3.7	20.1*	3.4
Group B Test Announced/Not Given *n*=35	23.2	4.2	13.9	3.8
Group C Test Not Announced/Not Given Control *n*=38	22.8	4.1	13.5	4.2
Overall *n*=110	22.8	4.0	15.9	3.8

*Means significantly higher at the .05 level

Table 2
Comparison of group means on the metals pretest via GLM procedure

Source	D.F.	Sum of Squares	Mean Square	*F*	*p*-value	Findings
Treatments	2	9.28	4.64	0.29	.748	n.s.
Error	107	1707.71	15.96			
Total	109	1716.99				

n.s. = not significant at the p<.05 level

The GLM procedure was then used to compare the 3 treatment groups on the means of the Delayed Retention Test scores. A significant difference was found among the total test means: $F(2, 107) = 34.69$, $p<.0001$ (see Table 3).

Follow-up comparisons were conducted via *t*-test (LSD) procedures in SAS. The results of the LSD comparisons are shown in Table 1. The critical value used was $t(107) = 1.98$, $p<.05$. The mean of the tested experimental group, Group A (Test), was significantly higher than either non-tested group, Group B (No Test) and Group C (Control). This was a clear demonstration of testing

effect—the act of taking the test helped students retain the information. The means of Groups B and C, however, did not differ significantly from each other even though Group B expected to be tested and graded on the material.

Table 3
Comparison of group means on the delayed retention test via GLM procedure

Source	D.F.	Sum of Squares	Mean Square	*F*	*p*-value	Findings
Treatments	2	1016.59	508.29	34.69	.0001	*
Error	107	1567.78	15.96			
Total	109	2584.37				

*Significant at the p<.05 level

The results of the survey administered to Group B are shown in Table 4. Only 11% of the students claimed that they would study if they did not expect a test. Nearly a third of the students reported that they consider themselves to be "test anxious beyond the level of most normal students." Other findings from the survey concerning which types of tests students prefer and which types they believe are most accurate for evaluation are also shown in Table 4.

Discussion

Three research questions were addressed by this study:

1. If delayed retention learning is the objective of instruction, does initial testing of the information aid retention learning? Within the constraints of this study, testing of instructional material did promote retention learning. This finding, a clear demonstration of testing effect, has been very consistent among several studies (Haynie 1990a, 1990b, 1991, 1994, 1995a; Nungester & Duchastel, 1982).

2. If delayed retention learning is the objective of instruction, does the anticipation of an upcoming test on the information aid retention learning? In some previous studies using a similar protocol the question was raised by reviewers whether it was the actual act of taking the test which aided retention learning or if the knowledge that a test was forthcoming motivated students to study more effectively. This was a central research question of one previous study (Haynie, 1990a) in which announcements of the intention to test were evaluated and shown not to be effective in promoting retention learning unless they were actually followed by tests or reviews. That finding was clearly repeated here because only the group which was actually tested (Group A) outscored the control group (Group C)—the students who expected a test but did not actually take the test (Group B) scored no better on retention than the control group which expected no test. Reviewers also criticized the previous studies because students in all groups had been told that their efforts would not count in their course grades, so they likely did not take a serious approach to their study of this unit. In this investigation, however, both of the experimental groups (A and B) did expect their scores on the immediate posttest to be counted in

determination of their course grades. Despite the fact that Group B expected a test to be given and expected it to be counted in their course grades, they were

Table 4
Results of survey on testing from Group B

	Yes		No	
Item Stem	#	%	#	%
I would study if there was no test expected	4	11.4	31	88.6
I am test anxious	11	31.4	24	68.6
I prefer this type of test:				
Take-Home	28	80.0	7	20.0
Multiple-Choice	31	88.6	4	11.4
True-False	9	25.7	26	74.3
Short Answer	15	42.9	20	57.1
Essay-Discussion	11	31.4	24	68.6
Matching	22	62.9	13	37.1
This type of test is more accurate:				
Take-Home	8	22.9	27	77.1
Multiple-Choice	18	51.4	17	48.6
True-False	3	8.6	32	91.4
Short-Answer	32	91.4	3	8.6
Essay-Discussion	29	82.9	6	17.1
Matching	12	34.3	23	65.7

n=35, Only Group B was surveyed

still outscored significantly by Group A. Since the metals pretest showed the groups to be of equal entering ability and everything else about the courses and treatments were the same, the only identifiable difference between these two groups was that Group A may have moved more information from short term to long term memory while they were engaged in the act of taking the test (testing effect), but Group B did not show these gains simply due to their supposed increased study or motivation—only actual testing brought about increased retention. This finding is consistent among the previous studies in this series, even though most of those studies did have the flaws mentioned above.

3. Do students study with greater effort when they expect a test than when they do not expect a test? In this study, the group which was told they would be tested but did not actually take the test did not show any gains in retention over the control group. This finding was consistent with a similar study by Haynie (1990a). However, in answer to an item on the survey, over 88% of the students reported that they would not study material unless they expect it to be reflected on a test—this would support the practice of administering regular preannounced tests to provide external motivation for students to study.

Other findings from the survey included: Students prefer take home, multiple-choice, and matching tests; but they acknowledge that take home and matching tests probably do not test knowledge very accurately. Additionally, about a third of the students feel test anxious, so there may be differential effects of the "pressure of a looming test" for these students vs. non test anxious students.

The conclusion here is that, in general, students do likely *study more earnestly* when they expect a test than if they do not, but maximum benefit in retention is gained only by having students anticipate and then actually take a test. The idle threat of an upcoming test did not result in increased retention for Group B in this study or in the earlier one (Haynie, 1990a), only Group A which was tested actually retained more knowledge after a delay of three weeks. Most readers will rightly assume that a large portion of the gains demonstrated here were due to simple testing effect (when a student takes the same test or an alternate form of a test a second time, the score is likely to increase). However, one-third of the information on the delayed retention test used in these studies is not reflected in any way on the initial posttests. The gains in retention were demonstrated by Group A in both the previously tested and the novel items of the delayed retention test. Group A scored 18 percentage points higher on the previously tested material and 23 percentage points higher on the novel material than did Group B, while Group B showed no gains over the control Group C in either subsection of the delayed retention test. Thus there is some evidence here that the gains may exceed those normally associated with simple testing effect. Therefore, this researcher concluded that being tested helps students to retain information while simply being warned of a test and expecting a grade does not.

Recommendations

Since testing consumes such a large amount of teacher and student time in the schools, it is important to learn as much as possible about the effects of tests on learning. It is important to maximize every aspect of the learning and evaluation process. The ability of teachers to develop and use tests effectively has been called into question recently, however, most research on testing has dealt with standardized tests. The whole process of producing, using, and evaluating classroom tests is in need of further research.

This study was limited to one educational setting. It used learning materials and tests designed to teach and evaluate a limited number of specified objectives concerning one body of subject matter. The sample used in this study may have been unique for unknown reasons. Though the present study did support findings of a study in a different setting, they must be replicated in numerous settings and via differing methods before they can be accepted. Therefore, studies similar in design which use different materials and are conducted with different populations will be needed to achieve more definite answers to these research questions. However, on the basis of this one study, it is recommended that: (a) when useful for evaluation purposes, classroom testing should continue to be employed due to its positive effect on retention learning, and (b) students should know in advance that they will be tested because of the effect this information may have on their study habits. The time devoted by teachers and

students to classroom testing apparently does have learning value in addition to its utility for evaluation purposes.

The value of tests in promoting retention learning has been demonstrated here and research questions about anticipation of tests have been addressed, however, there remain many more potential questions about classroom testing. The tests used in this study were carefully developed to resemble and perform similarly to teacher-made tests in most regards, however, there are still research questions which must be answered only on the basis of tests actually produced by teachers and for use in their natural settings.

References

Carter, K. (1984). Do teachers understand the principles for writing tests? *Journal of Teacher Education*, *35*(6), 57-60.

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

Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple choice test items: What are textbook authors telling teachers? *Journal of Educational Research*, *83*(5), 289-293.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement, No. 19* (pp.29-38). San Francisco: Jossey-Bass.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. *Educational Measurement: Issues and Practice*, *4*(1), 15-18.

Haynie, W. J. (1983). Student evaluation: The teacher's most difficult job. *Monograph Series of the Virginia Industrial Arts Teacher Education Council*, Monograph Number 11.

Haynie, W. J. (1990a). Effects of tests and anticipation of tests on learning via videotaped materials. *Journal of Industrial Teacher Education*, *27*(4), 18-30.

Haynie, W. J. (1990b). Anticipation of tests and open space laboratories as learning variables in technology education. In J. M. Smink (Ed.), *Proceedings of the 1990 North Carolina Council on Technology Teacher Education Winter Conference*. Camp Caraway, NC: NCCTTE.

Haynie, W. J. (1991). Effects of take-home and in-class tests on delayed retention learning acquired via individualized, self-paced instructional texts. *Journal of Industrial Teacher Education*, *28*(4), 52-63.

Haynie, W. J. (1992). Post hoc analysis of test items written by technology education teachers. *Journal of Technology Education*, *4*(1), 27-40.

Haynie, W. J. (1994). Effects of multiple-choice and short answer tests on delayed retention learning. *Journal of Technology Education*, *6*(1), 32-44.

Haynie, W. J. (1995a). In-class tests and posttest reviews: Effects on delayed-retention learning. *North Carolina Journal of Teacher Education*, *8*(1), 78-93.

Haynie, W. J. (1995b). An analysis of tests developed by technology teachers. Unpublished manuscript.

Hoepfl, M. C. (1994). Developing and evaluating multiple choice tests. *The technology Teacher, 53*(7), 25-26.

Herman, J., & Dorr-Bremme, D. W. (1982). *Assessing students: Teachers' routine practices and reasoning*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Mehrens, W. A. (1987). "Educational Tests: Blessing or Curse?" Unpublished manuscript, 1987.

Mehrens, W. A., & Lehmann, I. J. (1987). Using teacher-made measurement devices. *NASSP Bulletin, 71*(496), 36-44.

Newman, D. C., & Stallings, W. M. (1982). *Teacher Competency in Classroom Testing, Measurement Preparation, and Classroom Testing Practices*. Paper presented at the Annual Meeting of the National Council on measurement in Education, March. (In Mehrens & Lehmann, 1987)

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 447-474). New York: Macmillan.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*(1), 18-22.

Stiggins, R. J., & Bridgeford, N. J., (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5*(2), 5-17.