

Effects of Multiple-Choice and Short-Answer Tests on Delayed Retention Learning

W. J. Haynie, III¹

Abstract

This research investigated the value of short-answer in-class tests as learning aids. Undergraduate students ($n=187$) in 9 technology education classes were given information booklets concerning “high-tech” materials without additional instruction. The control group was not tested initially. Students in the experimental groups were either given a multiple-choice or a short-answer in-class test when they returned the booklets. All groups were tested for delayed retention three weeks later. The delayed retention test included subtests of previously tested and new information. Both short answer and multiple-choice tests were more effective than no test in promoting delayed retention learning. No difference was found between short-answer and multiple-choice tests as learning aids on the subtest of information which had not been tested on the initial tests, however, multiple-choice tests were more effective in promotion of retention learning of the information actually contained in the immediate posttests.

This study compared two types of teacher-made in-class tests (multiple-choice and short-answer) with a no test (control) condition to determine their relative effectiveness as aids to retention learning (that learning which is still retained weeks after the initial instruction and testing have occurred). The investigation involved instruction via self-paced texts, initial testing of learning, and delayed testing 3 weeks later. The delayed tests, which included both previously tested information and novel information that had not been previously tested, provided the experimental data for the study.

Background

The importance of testing in education makes it an important topic of continuing research. As technology education evolves to emphasize more cognitive learning, the time devoted to testing and the effects of testing will become increasingly important. Most of the research on testing which has been reported

¹William J. Haynie, III is an Associate Professor in the Department of Occupational Education, North Carolina State University, Raleigh, NC.

in recent years has concerned standardized tests (Bridgeford, Conklin, and Stiggins, 1986). Most of the evaluation done in schools, however, is done with teacher-made tests (Haynie, 1983, 1991, 1992; Herman & Dorr-Bremme, 1982; Mehrens, 1987; Mehrens & Lehmann, 1987; Newman & Stallings, 1982). The available findings on the quality of teacher-made tests cast some doubt on the ability of teachers to perform evaluation effectively (Burdin, 1982; Carter, 1984, Fleming & Chambers, 1983; Gullickson & Ellwein, 1985; Haynie, 1992; Stiggins & Bridgeford, 1985; Wiggins, 1993). Despite these problems, Mehrens and Lehmann (1987) point out the importance of teacher-made tests in the classroom to evaluate attainment of specific instructional objectives. Evaluation by teacher-made tests in schools is an important part of the educational system and a crucial area for research (Haynie, 1990a, 1990b, 1991, 1992; Mehrens & Lehmann, 1987; Wiggins, 1993).

One method of testing that has received little attention in the literature, however, which is popular in many educational settings, is the use of short-answer test items. Short-answer items are relatively easy to prepare (Haynie, 1983) and may be scored more quickly than essay items. They are not as objective as multiple-choice items because they sometimes do not give adequate information to evoke the desired response even from students who know the subject well. Despite this limitation, they may be useful on teacher-made tests because there is good evidence to suggest that many teachers are not capable of authoring truly clear and effective multiple-choice items (Haynie, 1983, 1992). Since many teachers do use short-answer items, their usefulness in promotion of retention learning is worthy of research.

Multiple-choice tests, take-home tests, and post-test reviews have all been shown to promote retention learning in previous studies (Haynie, 1990a, 1990b, 1991, in press; Nungester & Duchastel, 1982). However, announcements of an upcoming test did not have a positive effect on retention learning without a test actually being given. It appears that increased studying due to anticipation of a test did not result in better retention — only the act of taking the test increased retention (Haynie, 1990a). No studies were found that investigated the effects of short-answer tests on retention learning which is the thrust of this research. Research on the effects of tests on retention learning within the context of technology education classes and the value of the learning time they consume is limited to the studies cited above.

Purpose and Definition of Terms

The purpose of this study was to investigate the value of in-class multiple-choice and short-answer tests 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. A delay period of 3 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. (Dwyer, 1968; Dwyer, 1973; Duchastel, 1981; Nungester & Duchastel, 1982; Haynie, 1990a, 1990b, 1991, in press). The delayed retention test results were the only data analyzed in this investigation.

In addition to studying the relative gains in retention learning acquired by students while they take a test, an effort was made here to determine whether information which has been studied but which does not actually appear on the immediate posttest will be retained in addition to that material which is on the test. This study also examined whether multiple-choice and short-answer tests differ in their effectiveness for promoting retention of both tested and untested material. 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. Does initial testing by short-answer tests aid retention learning as effectively as initial testing with multiple-choice tests?
3. Will information which is not represented on initial testing be learned equally well by students tested via short-answer and multiple-choice tests?

Methodology

Population and Sample

Undergraduate students in 9 intact technology education classes were provided a booklet on new “high-tech” materials developed for space exploration. There were 187 students divided into three groups: (a) Multiple-choice test (Group A, $n=63$), (b) Short-answer test (Group B, $n=64$), and (c) No test (Control, Group C, $n=60$). All groups were from the Technology Education metals technology (TED 122) classes at North Carolina State University. Students were majors 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 this group. All groups were team taught by the researcher and his graduate assistant. Treatments were randomly assigned to each section. Random assignment, deletion of students majoring in Aerospace Engineering, and absences on testing dates resulted in final group sizes which were slightly unequal.

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. It was also pointed out, however, 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 regular subtests reflected the newly added subject matter.

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.

Six intact class sections over a two year period were randomly assigned to Group A or Group B (three each). Both groups were told to study the packet and that they would be asked to take a test on the material in class one week later. Students were told that participation was voluntary and the tests would not affect their grades. Both groups were requested to return the packets on the test date also. Students were told that the purpose of the study was to examine the types of answers given on the tests to see if there was a difference in the way questions were approached. They were also again told that the results would not affect their course grades and that participation was voluntary.

In order to obtain a control group, three randomly selected sections of students in the same course during the two semesters of the next year were given the same initial instructions. However, instead of announcing a test, the teacher told the students 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. It is acknowledged that these students who participated in the

study in a different year than the other two groups could have been a confounding variable; however, they did come from three intact class sections with the same teachers as the other groups. It was felt that this was the only way to insure that students truly believed they would not be tested on the material. If they had been mingled with the other two groups, they would have readily seen that some sort of testing had to occur sometime or there would be no data for the experiment from their group. This would have also spoiled the effectiveness of the evasive statements to the other two (experimental) groups that “types of answers” on the tests were the data of interest.

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. Students were told again that participation was voluntary. They were again asked to do their best and reminded that it did not affect their grades.

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 teacher provided all directions and neither teacher 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. Some 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. Equal numbers of Fall Semester and Spring Semester students were assigned to each group. Normal precautions were taken to assure a good learning and testing environment.

Instrumentation

The initial tests were parallel forms of a single 20 item test. The short-answer version was identical to the multiple-choice form except that there were no alternatives from which to choose responses and brief prose answers were required. Multiple-choice items had five response alternatives. The same information was reflected in both tests. It must be noted that, in general, short-answer tests tend to be used more often and appear to be more effective with lower level types of learning (Haynie, 1983), therefore, the information in this study was taught and tested primarily at the first three levels of the cognitive domain: (a) knowledge, (b) comprehension, and (c) 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 in-class test. These served as a subtest of previously tested information. The remaining ten items were similar in nature and difficulty to the oth-

ers, but they had not appeared in any form on either of the initial tests. These were interspersed throughout the test and they served as a subtest of new information. The subtest on new information was used to determine if retention learning gains were made during the study period or during the process of actually taking the tests — assuming that all of the information had been originally studied with relatively equal diligence, this information should be learned equally by all groups. If the type of test employed effected retention learning gains, then one of the tested groups would be expected to outperform the other one on the subtest of previously tested information.

The delayed retention test was developed and used in a previous study (Haynie, 1990). 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. Thorndike and Hagen (1977) assert that tests with reliability approaching .70 are within the range of usefulness for research studies.

Data Collection

Students were given initial instructions concerning the learning booklets and directed when to return the booklets and take the tests. The in-class immediate posttests were 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. on a microcomputer. 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 data consisted of the Delayed Retention Test means of the three groups. The means of the two subtest sections were then similarly analyzed by the one-way GLM procedure to detect differences in retention of previously tested and novel information. 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.

Findings

The means, standard deviations, and final sizes of the three groups on the delayed retention test (including the two subtests and the total scores) are pre-

sented in Table 1. The overall difficulty of the test battery and each subtest can be estimated by examining the grand means and the range of scores.

The grand mean of all participants was 16.63 with a range of 3 to 28 on the total 30 item test. The grand mean on the 20 item subtest of previously tested material was 12.32 with a range of 2 to 20, and the grand mean on the 10 item subtest of new information was 4.31 with a range of 0 to 9. No student scored 100% on the entire test and the grand means were close to 50% on each test, so the tests were relatively difficult. The grand means, however, were not used in any other analysis of the data.

Table 1

Means, Standard Deviations, and Sample Sizes

Treatment	Subtests					
	Total Test		Previously Tested		New Information	
	Mean	SD	Mean	SD	Mean	SD
Group A Multiple-Choice <i>n</i> =63	19.05	4.00	14.05	2.89	5.00	1.95
Group B Short-Answer <i>n</i> =64	16.86	4.72	12.48	3.28	4.38	2.03
Group C No Test Control <i>n</i> =60	13.85	4.57	10.33	3.14	3.52	1.97
Overall <i>n</i> =187	16.63	4.43	12.32	3.10	4.31	1.98

The GLM procedure was used to compare the 3 treatment groups (Group A, Multiple-choice Test; Group B, Short-answer Test; and Group C, Control) on the means of the total delayed retention test scores. A significant difference was found among the total test means: $F(2, 184) = 21.16, p < .0001, R\text{-Square} = .19$.

Following this significant finding, the GLM procedure was again employed to examine the means of each subtest. Significant differences were found among the means on the subtest of previously tested information, $F(2, 184) = 22.07, p < .0001, R\text{-Square} = .19$, and among the means on the subtest of new information, $F(2, 184) = 8.64, p < .0003, R\text{-Square} = .08$.

Followup comparisons were conducted via *t*-test (LSD) procedures in SAS. The results of the LSD comparisons are shown in Table 2. The critical value used was $t(184) = 1.97$, $p < .05$. In the total test scores and both subtests (previously tested and new information), the means of the two treatment groups which were previously tested, Group A (Multiple-choice Test) and Group B (Short-answer Test) were both significantly greater than the means of Group C (No Test Control).

Table 2*Contrasts of Group Means Via LSD Procedures*

	Groups and Means		
	Group C No Test	Group B Short-Answer	Group A Mult.-Choice
Total Test	13.85	16.86	19.05
Subtests:			
Previously Tested	10.33	12.48	14.05
New Information	3.52	<u>4.38</u>	<u>5.00</u>

Note. Means *not* underlined were significantly lower at the .05 level.

LSD followup comparisons also showed that Groups A and B were equal in their retention knowledge of the new information (10 item subtest of information which was not previously tested), but that Group A (Multiple-choice Test) outsourced Group B (Short-answer Test) significantly on the 20 item subtest of previously tested information and on the total test.

Discussion

The first of three research questions addressed by this study was: 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. Two types of tests were shown to be effective in supporting retention learning. The question could be raised 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 a 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. No attempt was made in this study to

separate the effectiveness of prior knowledge concerning upcoming tests from gains made while studying for and taking the tests.

The second research question was: Does initial testing by short-answer tests aid retention learning as effectively as initial testing with multiple-choice tests? The findings presented here provide evidence that multiple-choice tests promote retention learning more effectively than do short-answer tests. Both Group A and Group B scored significantly higher than the control (no test) group on the total test and both subtests. However, multiple-choice tests appear to be more effective in promoting retention learning than are short-answer tests as shown by the finding of significantly higher scores for Group A on the subtest of previously tested information. This may be because the correct answer to each item is provided along with the distractors in the multiple-choice items, but students had no cues to help them remember the answers, or even reconsider the issues, in the short-answer test items. Moving information from short term to long term memory is aided by rehearsal and, it appears that, multiple-choice test items are a more effective form of rehearsal than short-answer test items.

An alternate conclusion would be that the students who took the multiple-choice test performed better simply because the delayed retention test was in the same (multiple-choice) form. Further research should be conducted to examine this factor. The recommendation given here is to choose the type of test which is best suited to the educational objectives and trust that when it is used for evaluation, it will also aid in promotion of retention learning. However, if it is desirable to maximize the promotion of retention learning, then use of multiple-choice items on the test may be preferred over short-answer items. This does assume, however, that the multiple-choice items used will be good test items which are devoid of the errors in item development shown in previous research on test items authored by teachers (Haynie, 1992).

The final research question was: Will information which is not represented on initial testing be learned equally well by students tested via short-answer and multiple-choice tests? The delayed retention test used in this experiment contained a subtest of ten items interspersed throughout the test which had not appeared in any form on the initial tests. If the two types of test were equal in effectiveness, then both the subtests of new and of pretested information should have found no differences between the groups except for poorer performance by the control group. Alternatively, if one type of test were superior in promotion of retention learning, then one experimental group should outscore the other one on the subtest of previously tested material, but not on the subtest of new material.

Although the tests were short, there was no significant difference in the performance of the two previously tested groups on the ten item subtest of new information. Though both of these groups outscored the control (no test) group

significantly on this subtest of novel information, there was no difference between the two experimental groups. So, short-answer and multiple-choice tests were both equally effective in promotion of retention of information which was studied but which was not actually reflected in the test items. The conclusion here is, if short-answer tests are well suited to the type of learning objectives being tested from an evaluation viewpoint, then well developed short-answer tests should be equally effective in promoting retention learning of incidental information as multiple-choice tests.

Recommendations

Since testing requires considerable amounts of student and teacher time in the schools, it is important to maximize every aspect of the 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 teacher-made tests is in need of 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. 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; (b) both multiple-choice and short-answer tests promote retention learning, however, multiple-choice tests are more effective in this regard; (c) it appears that teachers who use short answer tests need not be overly concerned that students will only benefit from the learning of those specific facts represented on the test to the exclusion of information not represented because both short-answer and multiple-choice tests were shown to be equal in their ability to promote retention of material which was studied but not actually included on the test. So, if the instructor wishes to maximize the potential gains in retention made while students take a test, multiple-choice tests should be used, however, if short-answer tests are more appropriate for the evaluation situation present, their use will also benefit students' retention, although to a lesser degree. The ability of the individual instructor to develop good multiple-choice test items should be considered in making this decision.

Short-answer tests may have advantages of their own which make them useful in some situations because they do not force students to choose from a predetermined set of responses. Though some of the research examined in the review of literature for this study was critical of short-answer tests, the fact that

teachers have difficulty authoring effective multiple-choice items may make short-answer items a better choice for many situations. This study did not examine the effect of post test reviews when using short-answer tests. Such reviews have been shown to be helpful in promoting retention of information tested via multiple-choice tests. The effects (on retention) of post test reviews following short-answer tests should be addressed in future research.

Testing, pre test reviewing, post test reviewing, and occasional retesting require large amounts of learning time. As technology education moves away from the traditional “shop” setting of industrial arts and toward a more conceptually based curriculum, the teaching and testing of cognitive information increases in importance. More of the time of students and teachers will be consumed by testing and related activities such as pre and post test reviews. Technology teachers should understand how to make this time beneficial for learning as well as for evaluation. Technology teacher educators should help preservice and inservice teachers learn how to maximize the learning potential of time devoted to testing and reviews. The value of tests in promoting retention learning has been demonstrated here and two research questions about the types of tests to use for specific purposes within the context of technology education classes have been addressed, however, there remain many more potential questions about all sorts of teacher-made tests. 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. The process of pre-test reviewing, testing, and post test reviewing is too time consuming to be ignored. Continued research must be conducted to determine the best ways to test and review so as to meet the needs of evaluation and to maximize retention of important learning in technology education and in other disciplines.

References

- Burdin, J. L. (1982). Teacher certification. In H. E. Mitzel (Ed.), *Encyclopedia of education research* (5th ed.) New York: Free Press.
- Carter, K. (1984). Do teachers understand the principles for writing tests? *Journal of Teacher Education*, 35(6), 57-60.
- Denny, J. D., Paterson, G. R., & Feldhusen, J. F. (1964). Anxiety and achievement as functions of daily testing. *Journal of Educational Measurement*, 1, 143-147.
- Duchastel, P. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, 6, 217-226.
- Dwyer, F. M. (1968). Effect of visual stimuli on varied learning objectives. *Perceptual and motor Skills*, 27, 1067-1070.

- Dwyer, F. M. (1973). The relative effectiveness of two methods of presenting visualized instruction. *The Journal of Psychology*, 85, 297-300.
- Faw, H. W., & Waller, T. G. (1976). Mathemagenic behaviors and efficiency in learning from prose. *Review of Educational Research*, 46, 691-720.
- 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.
- Gay, L., & Gallagher, P., (1976). The comparative effectiveness of tests versus written exercises. *Journal of Educational Research*, 70, 59-61.
- 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. (in press). Effects of in-class tests and post-test reviews on delayed retention learning acquired via individualized, self-paced instructional texts. *Journal of Industrial Teacher Education*.
- 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)

- 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.
- Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and Evaluation in Psychology and Education*. New York: Wiley.
- Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, 75(3), 200-214.