Effects of Frequent Classroom Testing

ROBERT L. BANGERT-DROWNS
University at Albany
State University of New York

JAMES A. KULIK
CHEN-LIN C. KULIK
The University of Michigan

ABSTRACT **The educational effects of frequent classroom testing have been studied and discussed since the early part of this century. Testing advocates have suggested that more frequent classroom testing stimulates practice and review, gives students more opportunities for feedback on their work, and has a positive influence on student study time. Reviewers of relevant research and evaluation literature, however, have expressed uncertainty about whether such benefits are actually realized in classrooms. The present review distinguishes research on frequent classroom testing from research in two related areas, research on adjunct questions and research on mastery testing, and provides results from a meta-analysis of findings on frequency of classroom testing. The meta-analysis showed that students who took at least one test during a 15-week term scored about one half of a standard deviation higher on criterion examinations than did students who took no tests. Better criterion performance was associated with more frequent testing, but the amount of improvement in achievement diminished as the number of tests increased.**

U.S. teachers have not always been able to test students frequently on class material. Before the middle of the 19th century, writing materials were scarce in schools, and teachers had to use time-consuming oral recitations to check on student progress. When writing materials became more available during the late 19th century, teachers began using essay tests to evaluate students, but the difficulty of grading large numbers of essays limited the frequency with which the tests were given. Finally, during World War I, objective tests were developed that could be administered easily to large groups and then scored quickly and objectively. Frequent testing became an option for teachers in the United States.

The optimum amount of testing for a class also became a matter of controversy. Testing advocates argued that more frequent testing would increase instructional effectiveness and would encourage students to study and review more often. The advocates also contended that additional testing would provide opportunities for teachers to correct student errors, to reward good performance, and to give students a good indication of what they were expected to learn. But others noted that frequent testing could take time away from instruction. With a greater teacher emphasis on tests, some educators maintained that students might start directing their efforts toward performing well on tests rather than toward learning. Those educators also said that too frequent testing might inhibit integration of larger units of instructional materials and become tedious for students and, consequently, reduce their enthusiasm about learning.

Researchers have produced many studies that are relevant to this debate. Our purpose here is to review the research literature to draw some conclusions about the effects of frequent tests on students. We first describe and assess conclusions that have been drawn in earlier reviews of testing studies. We then assemble primary studies of frequent testing in classroom settings and report the results of a statistical analysis of the findings from those studies.

Earlier reviews related to frequent testing. The research that has been interpreted as showing the importance of frequent testing for learning comes from three major areas: (a) research on effects of adjunct questions on learning; (b) research on mastery testing; and (c) research on classroom quizzes and examinations.

Adjunct Questions

Adjunct questions are questions added to instructional text to influence learning from the text. Rothkopf (1966), who implemented pioneering studies on adjunct questions, referred to such questions as "test-like events." He thus implied that results from carefully controlled laboratory studies of adjunct questions in text might have relevance for other situations. In recent years, other researchers have made the same suggestion.

In a typical study on effects of adjunct questions, all the subjects are given an instructional text to study. The subjects in the experimental group receive adjunct questions along with the text, and they are told to answer the questions when they approach them. The subjects in the

Address correspondence to Robert Bangert-Drowns, ED 110, University at Albany/SUNY, 1400 Washington Avenue, Albany, NY 12222.

control group are given the text without the adjunct questions and are directed to just read the text. Experimental and control subjects take the same criterion examination after completing the experimental tasks. From the results on this examination, the experimenter is able to determine how much adjunct questions influence learning.

There are many variations on this basic design, however, and Hamaker (1986) has shown that at least three such variations influence experimental results. First, researchers have used questions at different cognitive levels in studies of adjunct questions. Some researchers have used lower order questions; some, higher order questions. Second, researchers have varied the placement of adjunct questions in a text. Some researchers have used prequestions, which precede the text; others have used postquestions, which follow it. And third, investigators varied the relation between adjunct questions and questions on a criterion examination. In his review, Hamaker defined three different relations: the criterion examination may be identical to adjunct questions (repeated questions); they may be different from but related to the adjunct questions (related questions); or they may be unrelated (unrelated questions).

Everyone agrees that repeated questions produce a large improvement on examination scores. The real question is whether adjunct questions affect performance on related and unrelated items on a criterion test. Hamaker (1986) has shown that effects of lower order, factual questions, are generally positive on related criterion items. Effects of higher order questions are also positive on related criterion items and are even stronger than those for factual questions. Effects on unrelated items are small for all types of questions. Factual prequestions may inhibit slightly the acquisition of unrelated information; higher order prequestions may have a slight positive effect. Hamaker has listed other factors that may influence the size of effect of adjunct questions, including text length, density of adjunct questions, adjunct question format, and criterion test format.

Can those results be generalized to classroom testing? Duchastel (1979) emphasized that findings from research on adjunct questions are suggestive of only what may happen in real classroom testing situations. Adjunct-question studies may be high in internal validity, but they are not high in external validity. Adjunct-question studies usually involve enough experimental control to allow for analytic study, but they take place in settings that are not like test situations. Results from such studies are not the best guide to what happens in classrooms when test frequency is increased.

Mastery Testing

Mastery tests are those tests that are used at frequent intervals during instruction to evaluate and guide student progress. Students who show mastery of the objectives covered on such tests are allowed to advance to new course material. Students who do not show mastery have their weaknesses diagnosed, receive corrective instruction, and are given new opportunities to show mastery. In most cases, the students are not allowed to advance to new material until they show mastery of the objectives covered in an earlier quiz.

Much research has shown that programs that incorporate mastery testing as one of their features often have positive effects on students. Keller's (1968) PSI, a mastery-oriented individualized teaching approach often used with college students, has an extraordinarily strong record of effectiveness (Kulik, Kulik, & Cohen, 1979; Kulik, Kulik, & Bangert-Drowns, 1990). Bloom's (1968) Learning-for-Mastery approach, a group-based teaching approach often used in elementary and secondary schools, also has a strong record of effectiveness (Block & Burns, 1976; Guskey & Gates, 1985; Kulik, Kulik, & Bangert-Drowns, 1990). Those studies do not focus on mastery testing, however; rather, they compare the effects of conventional teaching to effects of teaching systems that include mastery testing as one of their features.

Other research has shown that mastery testing may be the critical component in such teaching systems. A meta-analysis of 49 comparative studies showed that dropping mastery testing from Keller- and Bloom-type classes caused instructional effectiveness to drop substantially (Kulik & Kulik, 1986–87). The exact size of the drop depended on such factors as the stringency of the mastery criterion and the amount of experimental control used in the study.

Such results suggest that frequent testing can help students to learn. But the results do not show that frequent testing is sufficient in itself. Mastery testing is frequent testing combined with several other features. Students in mastery classes not only take more tests but also receive specific feedback on each test item and corrective instruction on the basis of their test performance. The mastery testing literature does not establish that simply increasing the number of tests in a class will increase student learning.

Frequent Classroom Testing

The most relevant research on the effects of frequent testing is applied research carried out in actual classrooms with real tests, and there is a long tradition of such research in educational psychology. Soon after the first objective tests were developed, researchers examined the tests for educational effects. Spitzer (1939), for example, cited early studies conducted in 1914 and 1922 to determine whether immediate tests of written recall improved retention of newly learned material. Jones (1923) carried out a program of experiments to determine whether brief examinations increased retention of information given in lectures. Turney (1931) and Kulp (1933) divided their students into two groups according to their performance on

an examination. In both studies, the group that had performed poorly on the examination was given weekly tests in addition to regular examinations: the group that originally performed well received only the regular examinations. Turney and Kulp reported that by the end of the course, the two groups performed equally well on the final examination.

The early researchers believed that tests stimulated rehearsal of newly learned material and thus inhibited the normal decay of memory. Their inquiries set the stage, however, for a long history of speculation about the educational potency of testing. Jones (1923) suggested that testing operates as active practice, strengthening the associations between stimuli and responses. McKeachie (1963) suspected that the ability of tests to provide knowledge of results explained their instructional potency. Rickards (1979) showed that even when they did not provide feedback, tests could stimulate processing that would facilitate retrieval of learned information. Mawhinney, Bostow, Laws, Blumenfeld, and Hopkins (1971) demonstrated that students who faced daily tests studied more consistently than did students who were tested less frequently.

Reviewers of educational literature have been less optimistic about the effects of frequent testing than those researchers and theorists were. Ross and Stanley (1954), for example, reported on results from 16 studies of frequent classroom testing. Eight of the studies found that frequent testing had positive effects on student achievement measured immediately after a course, but 2 studies found negative effects, and 6 studies reported either mixed or no effects. Proger and Mann (1973) conducted an extensive review of research on the effects of testing. They informally summarized evidence about frequent testing from 27 studies of varying quality. Proger and Mann concluded that the reports were too contradictory to demonstrate any improvement in learning with an increase in testing frequency.

No one has yet used quantitative review methods to summarize research findings on effects of frequency of classroom tests. Quantitative review methods would undoubtedly supplement the findings from the more conventional reviews. Such methods are especially helpful when findings in an area seem mixed and contradictory. Quantitative tools can often be used to find order in seeming confusion. Such tools are also useful when reviewers want to determine how large effects are and whether they vary systematically as a function of study features.

The method that we used to investigate the effects of frequency of classroom testing on students was study effect meta-analysis (Bangert-Drowns, 1986), a modification of the meta-analytic approach described by Glass, McGaw, and Smith (1981). Like Glass et al., study effect analysts (a) locate studies of an issue through replicable search procedures, (b) code the studies for salient features, (c) describe study outcomes on a common scale, and (d) use statistical methods to find relations betwen study features and study outcomes. Unlike Glass and his colleagues, however, study effect analysts treat each study as a single case in a statistical analysis. In study effect analysis, the sample size is the number of studies of a research topic, not the number of findings in the studies.

Method

Data Sources

To find studies on the effects of frequent testing, we performed computer searches of two library data bases: (a) ERIC, a data base on educational materials from the Educational Resources Information Center, consisting of the two files *Research in Education* and *Current Index to Journals in Education* and (b) *Comprehensive Dissertation Abstracts.* The empirical studies retrieved in those computer searches were the primary source of data for our analyses. A second source of data was a supplementary set of studies located by branching from bibliographies in the review articles retrieved by computer.

Search procedures yielded 40 studies that met four criteria of research relevance and methodological adequacy. The studies had to take place in real classrooms. Laboratory studies and research using paid volunteers were not considered relevant. Second, the studies had to compare groups that took different numbers of tests but otherwise received identical instruction. Third, tests given to both groups had to be conventional classroom tests. Studies of mastery testing, for example, or studies of adjunct questions were not used in our statistical analysis. Finally, studies that had serious methodological flaws, such as significant pretreatment differences between groups, were excluded from our analysis.

Study Features

Twelve variables were used to describe treatments, methodologies, settings, and publication histories of the studies. The 12 variables were chosen on the basis of an examination of variables used to describe study features in previous reviews and on a preliminary examination of dimensions of variation in the studies located for this analysis. Two coders independently coded each of the studies on each of the variables. The coders then jointly reviewed their coding forms and resolved any disagreements by reexamining studies in which coding was in dispute.

Three of the 12 variables described aspects of the testing procedures:

Number of tests for the experimental group. For semester-long studies, this variable includes the number of tests given to the frequently tested group. For studies of more or less than a semester's duration, we adjusted the

number to indicate the total number of tests that would have been administered during a 15-week term.

Number of tests for the control group. This variable was calculated in the same way as the number of tests for the experimental group.

Duration of treatment. Study duration was recorded in number of weeks.

Four variables were used to describe the experimental designs of the studies:

Subject assignment. Students were assigned to experimental and control groups either randomly or by nonrandom procedures.

Control for teacher effects. In studies with this control, the same instructor or instructors taught both the experimental and the control groups. In studies lacking this control, instructors were different for the experimental and the control groups.

Control for author bias in criterion examination. Studies controlling for bias in test authorship were those that used standardized, commercial tests as criterion measures. Studies lacking this control used as criterion measures either localy developed tests or a combination of locally developed and commercial tests.

Amount of statistical control in outcome measurement. Outcomes were measured on a posttest alone in some studies, whereas in other studies outcomes were measured as pretest-posttest differences or as covariance-adjusted posttest scores.

Three variables were used to describe features of the settings in which the evaluations were conducted:

Class level. Courses were at the precollege or college level.

Course content. The subject matter taught in the courses was either mathematics, science, or social sciences.

Student ability. Students were of high, mixed, or low ability.

Two variables were used to describe the publication histories of the studies:

Year of the report. The publication or release year of each study was recorded.

Source of the study. The three document types were (a) technical reports, including clearinghouse documents, papers presented at conventions, and so on; (b) dissertations; and (c) professional publications, including articles, scholarly books, and so forth.

Outcome Measures

The instructional outcome measured in 35 of the 40 studies was student learning, as indicated on achievement examinations given at the end of instruction. In addition to those examinations, two other outcomes were measured in the studies. The first outcome was overall performance when items were given on short, frequent quizzes in a course versus performance on the same items when given on longer, less frequent quizzes. Nine studies compared performance on items when given under the two conditions. A second outcome measured in the studies was change in student attitude toward instructional method. Four studies contained results on this outcome measure.

For statistical analysis, outcomes had to be expressed on a common scale of measurement. We coded each outcome as an *effect size*, defined as the difference between the mean scores of two groups divided by the standard deviation of the control group. For most studies, effect sizes could be calculated directly from reported means and standard deviations. For some studies, however, effect sizes had to be retrieved from t and F ratios. Formulas used in estimating effect sizes from such statistics were those given by Glass, Cohen, Smith, & Filby, (1981).

In some studies, more than one value was available for use in the numerator of the formula for calculating effect size, and more than one value was available for the denominator. In such cases, we used as the numerator in the effect-size formula the difference that was least likely to be affected by individual differences among subjects and by other irrelevant factors. Therefore, we used covariance-adjusted differences rather than raw-score differences and differences in gains rather than differences on posttests alone. In addition, some reports contained several measures of variation that might be considered for use as the denominator in the formula for calculating effect size. We used the measure that provided the best estimate of the unrestricted population variation in the criterion variable.

Unit of Statistical Analysis

Some studies reported more than one finding for a given outcome area. Such findings sometimes resulted from the use of more than one experimental or control group in a single study, and they sometimes resulted from the use of several subscales and subgroups to measure a single outcome.

Representing a single outcome in a single study by several effect sizes violates the assumption of independence necessary for many statistical tests and also gives undue weight to studies with multiple groups and scales. Therefore, we calculated only one effect size for each outcome area of each study. Three rules helped us to decide which effect size best represented the study's findings. First, when results from both a true experimental comparison and a quasi-experiment were available from the same study, we recorded results of the true experiment. Second, when results from high-, intermediate-, and low-test frequencies were available in a single study, we used results from the high and low frequencies to calculate the effect size. Third, in all other cases, we used total scores and total group results rather than subscore and subgroup results in calculating effect sizes.

Results

Because 35 of the 40 studies in the pool investigated the effects of frequent classroom testing on criterion examination performance, we implemented a complete statistical analysis of results in that area. The analysis covered both average effects and the relationship between study effects and study features. We carried out less complete statistical analyses of other outcome areas because of the limited number of studies in those areas.

Examination Performance

Twenty-nine of the 35 studies with results from criterion examinations found positive effects from frequent testing, and 6 studies found negative effects. Thirteen of the 29 studies with positive findings reported that the difference in posttest achievement between experimental and control groups was statistically significant. Only one of the negative reports was statistically significant. Those box-score results indicate that frequent classroom testing is beneficial to student achievement (see Table 1).

The index of effect size provided a more precise measure of the strength of the treatment effects. The average of the 35 effect sizes was 0.23. That is, the average effect of frequent testing was to raise achievement scores by 0.23 standard deviations. The standard error of the mean

Table 1.—Major Features and Achievement Effect Sizes in 35 Studies of Frequent Classroom Testing

Study	Place	Class level	Course content	Duration in weeks	No. of tests		Effect size
					X group	C group	
Curo (1963)	Indiana	11th	Social science	6	25	2	0.10
Deputy (1929)	State University of New York	College	Philosophy	6	12	0	0.96
Dineen, Taylor, & Stephens (1989)	Nebraska	High school	Mathematics	15	75	15	0.17
Fitch, Drucker, & Norton (1951)	Purdue University	College	Government	15	15	4	0.26
Fulkerson & Martin (1981)	Western Illinois University	College	Psychology	12	8	4	0.07
Gable (1936)	Maryland	High school	Science	7	21	5	−0.80
Keys (1934)	University of California	College	Psychology	15	8	2	−0.01
Kirkpatrick (1934)	Iowa	High school	Science	18	20	2	0.31
Laidlaw (1963)	Fairleigh Dickinson University	College	Psychology	16	16	4	−0.08
Lindenberg (1984)	Illinois Community Colleges	College	Accounting	17	12	2	0.01
Mach (1963)	California State Polytechnic College	College	Mathematics	12	29	1	0.15[a]
Maloney & Ruch (1929)	California	9–11	Reading	10	5	0	0.59
Marso (1970)	University of Nebraska	College	Psychology	15	6	3	0.14
Monk & Stallings (1971)	University of Illinois	College	Geography	15	10	6	0.07
Mudgett (1956)	University of Minnesota	College	Engineering	12	36	2	0.26
Nation, Knight, Lamberth, & Dyck (1974)	University of Oklahoma	College	Psychology	8	8	1	−0.22
Negin (1981)	Marquette University	College	Law	15	3	0	0.70
Noll (1939)	Rhode Island State College	College	Psychology	15	5	1	−0.27
Nystrom (1969)	California Junior College	College	Mathematics	15	50	5	0.31
Olsen, Weber, & & Dorner (1968)	University of Illinois	College	Veterinary medicine	15	10	3	0.14
Palmer (1974)	Davison College	College	Psychology	10	6	0	0.55
Pikunas & Mazzota (1965)	Michigan	12th	Science	6	6	0	0.71
Pratt (1970)	Arizona	High school	Social science	9	11	0	0.19
Robinson (1972)	Brigham Young University	College	Psychology	4	3	0	0.10
Rievman (1974)	Florida Atlantic University	College	Psychology	16	10	2	0.34
Ross & Henry (1939)	Iowa State University	College	Psychology	12	10	1	0.06
Selakovich (1962)	West Texas State College	College	Government	15	15	3	0.08
Shapiro (1973)	New York Community College	College	Business	15	10	3	0.26
Standlee & Popham (1960)	Indiana University	College	Psychology	15	14	1	0.26
Stephens (1986)	University of Nebraska	College	Statistics	5	5	0	0.67
Townsend & Wheatley (1975)	California State Polytechnic College	College	Mathematics	12	49	1	0.54[a]
Ward (1984)	Western Illinois University	College	Statistics	15	13	2	−0.15[a]
Wiggins (1968)	University of North Carolina	College	Sociology	5	4	0	0.79
Wilkins (1979)	Louisiana Community College	College	Psychology	12	11	0	0.30
Williams & Lawrence (1974)	Western Michigan University	College	Physiology	4	3	0	0.34

[a]Estimated on the basis of direction and statistical significance of reported values.

was 0.06. That effect was significant by conventional statistical standards, $t(34) = 3.94$, $p < .001$.

One can examine the standard normal curve to get a clearer notion of the meaning of this effect size. Fifty-nine percent of the standard normal curve fell below a z score of 0.23, indicating that the average student from frequently tested classes performed at the 59th percentile on a posttest; the typical student taught with less frequent testing performed at the 50th percentile on the same posttest. In other words, the average student who was frequently tested outperformed 59% of the students who were not frequently tested.

Examination Performance and Study Features

The 35 figures showed much variation in their outcomes (Figure 1). Deputy (1929) reported the largest effect, 0.96 standard deviations. At the other extreme, Gable (1936) reported a decrease of −0.80 standard deviations in examination scores after frequent testing was implemented in a course. Such variation in study outcomes may be unsystematic and may be produced by unique features of specific studies. But the possibility also exists that the variation in study results is systematic. To test this possibility, we carried out further analyses (Table 2).

Testing frequency in the control condition was the most important predictor of effect size. That frequency varied from no tests for the control group (in 11 studies) to 15 tests (in 1 study). Effect sizes were almost invariably moderately high when the frequently tested group was compared with a control group that received no tests. In 11 such studies, average effect size was 0.54. When control students received one test, however, effect sizes dropped precipitously to an average of 0.15. Thus, taking one test during a 15-week term seemed to provide almost as much preparation for a criterion examination as did higher test frequencies. Control-group students seemed to be greatly disadvantaged only when they took no tests before the criterion examination.

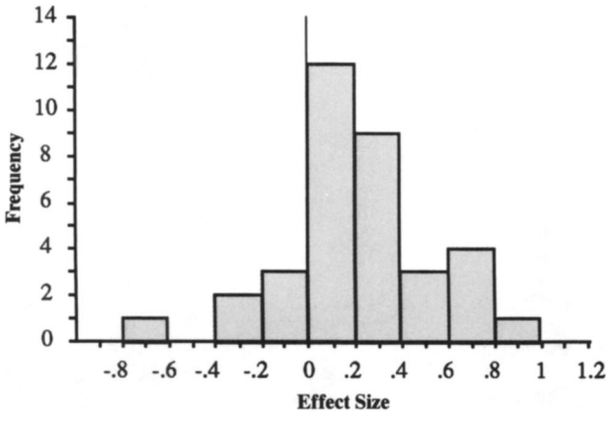

Figure 1. Histogram Showing the Distribution of 35 Achievement Effect Sizes From Studies of Frequent Classroom Testing

Table 2.—Means and Standard Errors of Achievement Effect Sizes for 35 Studies on Frequent Classroom Testing, by Study Feature

Study feature	*N*	*M*	*SE*
Adjusted number of tests for X Group[a]			
1 to 10	11	0.24	0.09
11 to 20	16	0.23	0.07
21 or more	8	0.21	0.18
Adjusted number of tests for C groups[a]*			
None	11	0.54	0.08
One	5	0.15	0.13
Two or more	19	0.07	0.06
Duration of treatment			
8 weeks or less	9	0.29	0.19
9 to 12 weeks	9	0.30	0.07
13 weeks or more	17	0.15	0.05
Subject assignment			
Random	6	0.36	0.10
Nonrandom	29	0.20	0.07
Control for instructor effect			
Same	29	0.18	0.06
Different	5	0.39	0.17
Control for author bias in criterion examination			
Commercial	7	0.16	0.19
Local and other	28	0.24	0.06
Statistical control in outcome measures			
Postscores only	20	0.28	0.08
Control for covariate	15	0.15	0.08
Class level			
Precollege	7	0.18	0.19
College	28	0.24	0.06
Course content			
Mathematics	6	0.28	0.12
Science	6	0.13	0.21
Social science	17	0.16	0.06
Others	6	0.46	0.14
Year of report			
Up to 1960	10	0.16	0.15
1961 to 1970	10	0.25	0.09
1971 to 1980	9	0.25	0.08
1981 or later	6	0.25	0.15
Source of study			
Unpublished	3	0.41	0.20
Dissertation	10	0.18	0.04
Published	22	0.22	0.09

[a]Number of tests was adjusted for each study to estimate the number of tests that would be given to students over a 15-week interval.
*$p < .001$.

Although the number of tests given to the control group was a good predictor of effect size, the number of tests given to the frequently tested group seemed unrelated to effect size. The raw correlation between effect size and the number of tests given to the frequently tested group was small and nonsignificant, −.06. The result was unexpected, and indicated that performance on criterion examinations does not increase as the number of tests increases in frequently tested classes. Thus, differences in the number of tests received by experimental groups in those studies—differences at the higher end of

the spectrum of test frequency—may be unimportant for student learning.

One source of evidence, however, suggested that frequency of testing in the experimental group might matter. The evidence came from eight studies that compared high, intermediate, and low frequencies of testing (Table 3). Such within-study comparisons are important because they are better controlled than between-study comparisons. In seven of the eight studies, the high-frequency group scored higher on criterion examinations than did the intermediate-frequency group. The results are even clearer when expressed as effect sizes. The average effect size for the high-frequency groups was 0.49; the average effect size for intermediate-frequency groups was 0.23. The difference in effect sizes for high- and intermediate-testing frequencies was statistically significant, $t(7) = 2.94$, $p < .05$.

Data from the eight studies hint at the kind of relationship that one might find between frequency of testing and effect size. On average, the intermediate-frequency groups received 7 tests during instruction (adjusted for a 15-week semester), whereas the higher frequency groups received 23 tests. Thus, a threefold increase in test frequency (from 7 to 23 tests) resulted in only a doubling of effect size (from 0.23 to 0.49 standard deviations). That finding suggests that increasing test frequency may regularly improve postinstruction achievement, but the improvement diminishes as test frequency increases—a case of diminishing returns.

A Model Relating Test Frequency and Classroom Learning

In a typical meta-analysis, a researcher studies the effects of fairly uniform experimental and control treatments. In such cases, the average effect size provides a good indicator of the overall effect of the treatment. In research on test frequency, however, experimental and control treatments vary in degree from study to study. Frequent testing may mean 3 tests a term in one study and 75 tests a term in another. Infrequent testing may mean no tests in one study and 15 tests a term in another. In similar cases, average effect sizes can give a misleading impression of treatment effects.

Glass et al. (1982) examined this type of problem in their analysis of results from studies of class size. They observed that small class and large class were relative terms, and what was a small class in one study might be a large class in another. They reasoned that if related at all, class size and achievement would be related in an exponential or geometric fashion. One pupil with 1 teacher would learn some amount; 2 pupils would learn less; 3 pupils would learn still less, and so on; and the drop from 1 to 2 pupils would be expected to be larger than the drop from 2 to 3, which, in turn, would probably be larger than the drop from 3 to 4, and so on. A logarithmic curve describes the relationship well, and Glass and his colleagues developed a general method for fitting that type of curve to results from treatments that vary in degree from study to study.

Data on testing frequency seems formally similar to Glass's class-size data. Logical analysis suggested that the relationship between test frequency and student learning should be monotonic, with larger numbers of tests generally associated with greater learning. The analysis also suggested that the relationship between test frequency and learning will, if anything, be exponential or geometric. That is, increasing the number of tests from 0 to 1 should make more of a difference in learning than should increasing the number of tests from 5 to 6 or from 10 to 11. Empirical results seemed to support our analysis, so we investigated the possibility of finding an exponential function to describe our data.

To fit an exponential model to our data, we first transformed test frequencies using various exponents: 1, 1/2, 1/3, 1/4, and so on. We then subtracted the transformed number of tests for the control group from the transformed number of tests for the experimental group, because difference in testing frequency should be of key im-

Table 3.—Effect Sizes in 8 Studies Using High, Intermediate, and Low Frequency of Testing

Study	Number of tests			Effect size	
	High-frequency condition	Intermediate-frequency condition	Low-frequency condition	High vs. low frequency of testing	Intermediate vs. low frequency of testing
Deputy (1929)	12	6	0	0.96	0.26
Mudgett (1956)	36	12	2	0.26	0.06
Negin (1981)	3	1	0	0.70	0.36
Palmer (1974)	6	3	0	0.55	0.18
Reivman (1973)	10	4	2	0.34	0.26
Shapiro (1973)	10	5	3	0.26	0.19
Townsend & Wheatley (1975)	48	3	1	0.54[a]	0.15[a]
Wilkins (1979)	11	3	0	0.30	0.38

[a]Estimated on the basis of direction and statistical significance of reported values.

portance. Finally, we correlated the differences with effect size. The raw difference between the number of tests of experimental and control groups correlated .02 with effect size; difference between square roots of test frequencies correlated .33 with effect size; difference between cube roots correlated .56; and difference between fourth roots correlated .63 (see Figure 2).

Figure 2 is based on a regression equation in which effect sizes are predicted from differences between the fourth roots of the test frequencies for experimental and control groups. The figure gives expected effect size as a function of the number of tests taken by the experimental group when the control group takes no tests. The figure shows, for example, that increasing the number of tests given per term from zero to two would raise performance on a criterion examination by 0.41 standard deviations. The figure can also be used to determine expected gains from other changes in test frequency. Thus, increasing the number of tests per term from two (with an effect size of 0.41) to four tests (with an effect size of 0.49) should increase student achievement by 0.08 (or 0.49–0.41) standard deviations. Perhaps the most important implication that can be drawn from this figure is that increasingly smaller gains would be achieved by increasing test frequencies above those that are already common in schools.

At first glance, the within-study evidence from eight studies comparing low-, intermediate-, and high-test frequencies (Table 3) seemed to yield results different from the between-study findings of 35 studies. The regression equation helped to demonstrate that the within-study evidence was consistent with the between-study evidence. In the eight studies listed in Table 3, the average number of tests (adjusted for a 15-week term) was 1, 7, and 23 in the low-, intermediate-, and high-test frequency groups, respectively. From the regression equation, one would predict that the effect sizes associated with those test frequencies would be 0.34, 0.56, and 0.74, respectively. Therefore, one would predict that intermediate-frequency group (increasing from 1 to 7 tests) would improve its criterion performance by 0.21 standard deviations (0.56–0.34) and the high-frequency group (increasing from 1 to 23 tests) to improve criterion performance by 0.40 standard deviations (0.74–0.34). The average effect size for the intermediate-frequency groups was 0.23, and 0.49 for the high-frequency groups.

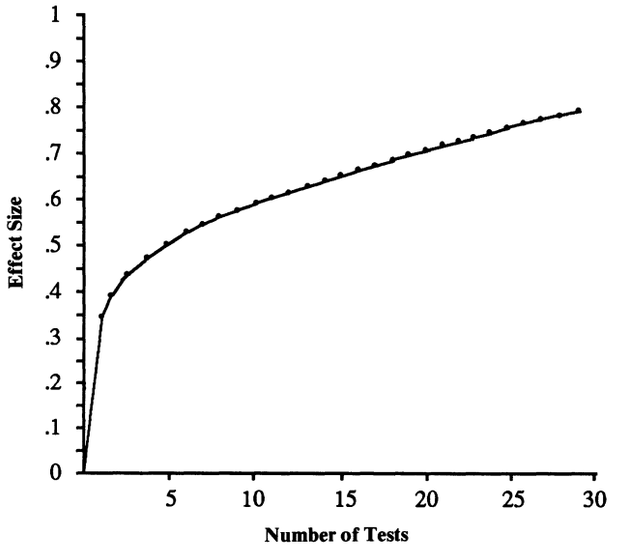

Figure 2. Expected Effect Size When Test Frequency Is Increased From No Tests During a 15-Week Term

Table 4.—Attitude Effect Sizes in Four Studies on Frequent Classroom Testing

Study	Effect size
Fulkerson & Martin (1981)	0.80
Gaynor & Milham (1976)	0.49
Nystrom (1968)	0.31
Shapiro (1972)	0.74

Attitudes Toward Instruction

Four studies measured students' attitudes toward instruction after they were exposed to differing testing conditions (Table 4). In all four studies, frequently tested students rated their classes more favorably than did students who were less frequently tested. The average of effect sizes drawn from the four studies was large (0.59 standard deviations).

Performance on Differentially Distributed Test Items

Fourteen studies were found in which experimental and control students were tested with the same items during instruction. In the experimental classes, the items were distributed in short, frequent quizzes; in the control classes, the identical items were given in longer, less frequent tests. The studies, therefore held constant the number of test items given during instruction and examined whether differential distribution of the items affected academic performance.

Nine studies reported the performance of students on the test items given during instruction. In the typical study, the experimental group outperformed the control group by 0.57 standard deviations. The students performed better on short quizzes on short units of instruction than on longer quizzes over longer units of instruction (Table 5).

Does improved performance on test items distributed more extensively during instruction translate into better posttest performance? Apparently it does not. Nine studies with differential distribution of identical items reported the performance of students on posttests (Fulkerson & Martin, 1981; Gable, 1936; Keys, 1934; Laidlaw, 1963; Lindenberg, 1984; Marso, 1970; Monk & Stallings, 1971; Rievman, 1973; Shapiro, 1973). The average effect

Table 5.—Features and Achievement Effect Sizes in Studies Comparing Different Distributions of Quiz Items

Study	Place	Number of tests		Effect size
		X group	C group	
Badia, Harsh, & Stutts (1978)	Bowling Green State University	9	3	0.25
Dustin (1971)	SUNY, Plattsburg	4	1	0.62
Fulkerson & Martin (1981)	Western Illinois University	8	4	0.60
Gaynor & Milham (1976)	University of Houston	12	2	0.21
Keys (1934)	University of California	8	2	0.53
McDaris (1985)	University of Oklahoma	3	1	1.28
Monk & Stallings (1971)	University of Illinois	8	4	0.22
Stephens (1977)	University of Nebraska	16	2	0.38
Rievman (1973)	Florida Atlantic University	10	2	1.02

size comparing the more frequently tested group to the less frequently tested group was zero.

Discussion

Since the second decade of this century, researchers have speculated about the effects of frequent testing on classroom learning. Some early researchers (Jones, 1923) expected great benefits from classroom testing, but others (Noll, 1939) complained about possible negative effects from too much testing. Reviewers who have examined the research results have been unable to reconcile apparent contradictions in findings (Proger & Mann, 1973; Ross & Stanley, 1954), and they have not reached definite conclusions about effects of frequent testing.

Based on research in related fields, one expects positive effects from frequent classroom testing. Research on adjunct questions has consistently (Rickards, 1979) shown that dividing texts into small units with "test-like events" improves student achievement. Research has also shown that mastery testing, when used as a diagnostic tool and followed with remedial help, also improves classroom learning (Kulik & Kulik, 1986–87).

The conditions that characterize ordinary classroom testing are much different from those that prevail in adjunct question research and in studies of mastery testing. In research on adjunct questions, for example, presentations of text material are typically short and simple. In the classroom, on the other hand, the information to be learned is usually greater in amount and complexity and is presented in a variety of ways, such as text, lecture, audiovisuals, and exercises. The students are generally encouraged to review material before a test, and the review may even be led by the teacher. Some students may rely on cramming to compensate for inattentiveness to instruction. Ordinary classroom tests are often used without feedback and correctives as extensive as that used with mastery testing. With ordinary classroom tests, the students are usually aware that their test performance is a *one-time* event that contributes to the student's academic record.

Our meta-analysis showed that the use of classroom testing does increase performance on criterion measures of achievement, but at a diminishing rate of return. When tested groups were compared with groups who received no tests, the tested groups typically scored about one half standard deviation higher on a criterion examination than did the untested students. However, few teachers could improve their instructional effectiveness that much simply by adding more tests to their courses. Results from our regression analysis suggest, for example, that a teacher who gives two tests during a term would increase examination scores by only 0.08 standard deviations after doubling test frequency. The effects on student learning from increasing test frequency would be most notable for those few teachers who give no tests except a final in a course. According to our regression analysis, such teachers would notice an increase in examination scores of 0.34 standard deviations if they added only one test to their course plan. Gains are incrementally smaller with each test added to the course.

We also found that when two groups answered identical test items, superior performance was obtained from students who answered the questions on a large number of short tests rather than on a small number of long tests. The total number of items answered correctly was 0.57 standard deviations higher for students who took short tests. Studies of this sort, however, cannot be used to draw conclusions about effects of frequent testing because the studies do not investigate performance on a common criterion examination given to both experimental and control groups under the same conditions. Available evidence suggests that differential distribution of test items during instruction has no effect on criterion test performance.

Finally, this meta-analysis shows that teachers can improve the affective outcomes of instruction by testing students more often. Four studies measured students' attitude toward instruction following programs of varying test frequency. The frequent testing condition had the effect of making students' attitudes more positive by 0.59 standard deviations. That is, students in those studies had

a more favorable opinion of their instruction when they were tested frequently. Increasing the frequency of tests may be a way of creating a more positive atmosphere in the classroom.

REFERENCES

Badia, P., Harsh, J., & Stutts, C. (1978). An assessment of methods of instruction and measures of ability. *Journal of Personalized Instruction, 3*, 69–75.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99*, 388–399.

Block, J. H., & Burns, R. B. (1976). Mastery learning. In L. S. Shulman (Ed.), *Review of research in education* (Vol. 4). Itasca, IL: F. E. Peacock.

Bloom, B. S. (1968, May). Mastery learning. *Evaluation Comment, 1*(2). Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation of Instructional Programs.

Curo, D. M. (1963). An investigation of the influence of daily pre-class testing on achievement in high school American history classes. *Dissertation Abstracts International, 24*, 5236. (University Microfilms No. 64–4574)

Deputy, E. C. (1929). Knowledge of success as motivating influence in college work. *Journal of Educational Research, 20*, 327–334.

Duchastel, P. C. (1979). *Adjunct questions effects and experimental constraints. Occasional paper 1.* Bryn Mawr, PA: American College. (ERIC Document Reproduction Service No. ED 216 312)

Dustin, D. (1971). Some effects of exam frequency. *The Psychological Record, 21*, 409–414.

Fitch, M. L., Drucker, A. J., & Norton, J. A., Jr. (1951). Frequent testing as a motivating factor in large lecture classes. *Journal of Educational Psychology, 42*, 1–20.

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

Gable, F. (1936). *The effect of two contrasting forms of testing upon learning* (Studies in Education Series, No. 25). Baltimore: Johns Hopkins University.

Gaynor, J., & Willham, J. (1976). Student performance and evaluation under variant teaching and testing methods in a large college course. *Journal of Educational Psychology, 68*, 312–317.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size: Research and policy.* Beverly Hills: Sage.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage Publications.

Guskey, T. R., & Gates, S. L. (1985, April). *A synthesis of research on group-based mastery learning program.* Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 262 088)

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56*, 212–242.

Johnson, K. R., & Ruskin, R. S. (1977). *Behavioral instruction: An evaluative review.* Washington, DC: American Psychological Association.

Jones, H. E. (1923). Experimental studies of college teaching: The effect of examination on permanence of learning. *Archives of Psychology, 10*, 1–70.

Keller, F. S. (1968). "Goodbye, teacher . . . " *Journal of Applied Behavioral Analysis, 1*, 79–89.

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology, 25*, 427–436.

Kirkpatrick, J. E. (1934). *The motivating effect of a specific type of testing program.* (Doctoral dissertation, Iowa University, 1933). *University of Iowa Studies in Education, 9*, 41–68.

Kulik, C.-L. C., & Kulik, J. A. (1986–87). Mastery testing and student learning: A meta-analysis. *Journal of Educational Technology Systems, 15*, 325–345.

Kulik, C.-L., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastrey learning programs: A meta-analysis. *Review of Educational Research, 60*, 265–299.

Kulik, J. A., Kulik, C.-L., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34*, 307–318.

Kulp, D. H., II. (1933). Weekly tests for graduate students? *School and Society, 38*, 157–159.

Laidlaw, W. J. (1963). The effects of frequent tests on achievement, retention, and transfer, and test behavior. *Dissertation Abstracts International, 24*, 5197. (University Microfilms No. 64–4322)

Lindenberg, T. S. (1984). The effect of test frequency on achievement in the first principles of accounting course. *Dissertation Abstracts International, 45*, 2736A. (University Microfilms No. DA8426700)

Mach, G. R., Jr. (1963). A comparative study of student performance in an intermediate calculus class as a result of different evaluation programs. *Dissertation Abstracts International, 24*, 5248. (University Microfilms No. 64–5745)

Maloney, E. L., & Ruch, G. M. (1929). The use of objective tests in teaching as illustrated by grammar. *School Review, 37*, 62–66.

Marso, R. N. (1970). Classroom testing procedures, test anxiety, and achievement. *Journal of Experimental Education, 38*, 54–58.

Mawhinney, V. T., Bostrow, D. E., Laws, D. R., Blumenfeld, G. J., & Hopkins, B. L. (1971). A comparison of students studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis, 4*, 257–264.

McDaris, M. A. (1985, May). *Testing frequency revisited: A pilot study.* Paper presented at the annual meeting of the International Communication Association, Honolulu. (ERIC Document Reproduction Service No. ED 265 175)

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

McKeachie, W. J. (1963). Research on teaching at the college and university level. In N. L. Gage (Ed.), *Handbook of Research on Teaching.* Chicago: Rand McNally.

Monk, J. J., & Stallings, W. M. (1971). Another look at the relationship between frequency of testing and learning. *Science Education, 55*, 183–188.

Mudgett, A. G. (1956). The effects of periodic testing on learning and retention in engineering drawing. *Dissertation Abstracts International, 16*, 2351–2352. (University Microfilms No. 56–3744)

Nation, J. R., Knight, J. M., Lamberth, J., & Dyck, D. G. (1974). Programmed student achievement: A test of the avoidance hypothesis. *The Journal of Experimental Education, 42*, 57–61.

Negin, G. A. (1981). The effects of test frequency in a first-year torts course. *Journal of Legal Education, 31*, 673–676.

Noll, V. H. (1939). The effect of written tests upon achievement in college classes: An experiment and a summary of evidence. *Journal of Educational Research, 32*, 345–358.

Nystrom, N. K. (1969). An experimental study to compare the relative effects of two methods of instruction on learning of intermediate algebra. *Dissertation Abstracts International, 29*, 3532A-3533A.

Olsen, R. E., Weber, L. J., & Dorner, J. L. (1968). Quizzes as teaching aids. *Journal of Medical Education, 43*, 941–942.

Palmer, E. L. (1974). Frequency of tests and general subject-area mastery. *Psychological Reports, 35*, 422.

Pikunas, J., & Mazzota, D. (1965). The effects of weekly testing in the teaching of science. *Science Education, 49*, 373–376.

Pratt, F. H. (1970). The effect of frequent testing in American History upon student achievement and attitudes. *Dissertation Abstracts International, 31*, 80A. (University Microfilms No. 70–11,892)

Proger, B. B., & Mann, L. (1973). *An historical review of theoretical and experimental literature on the teaching values of informal (nonstandardized), teacher-made achievement tests: 1913–1968.* Blue Bell, PA: Montgomery County Intermediate Unit 23. (ERIC Document Reproduction Service No. ED 084 292)

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

Rievman, S. P. (1973). Optimal frequency of testing as a function of ability level and reinforcement history. *Dissertation Abstracts International, 34*, 7054A. (University Microfilms No. 74–11, 605)

Robinson, P. (1972). *Contingent systems of instruction.* Paper presented at the Rocky Mountain Psychological Association Convention. (ERIC Document Reproduction Service No. ED 069 704)

Ross, C. C., & Henry, L. K. (1939). The relation between frequency of testing and progress in learning psychology. *Journal of Educational Psychology, 30*, 604–611.

Ross, C. C., & Stanley, J. C. (1955). *Measurements in Today's Schools.*

New York: Prentice-Hall.
Rothkopf, E. Z. (1966). Learning from written instruction material: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3*, 241–249.
Selakovich, D. (1962). An experiment attempting to determine the effectiveness of frequent testing as an aid to learning in beginning college courses in American Government. *The Journal of Educational Research, 5*, 178–180.
Shapiro, S. L. (1973). An experimental study of the effects of frequency of testing procedures on students in a business organization and management course in a community college with an open admissions policy. *Dissertation Abstracts International, 35*, 4207A–4208. (University Microfilms No. 74–19,774)
Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641–656.
Standlee, L. S., & Popham, W. J. (1960). Quizzes' contribution to learning. *Journal of Educational Psychology, 51*, 322–325.
Stephens, L. J. (1977). The effect of the class evaluation method on learning in certain mathematics courses. *International Journal of Mathematical Education in Science and Technology, 8*, 477–479.
Stephens, L. J. (1986). The effect of frequent testing upon performance in mathematics courses. *International Journal of Mathematical Education in Science and Technology, 18*, 611–613.
Townsend, N. R., & Wheatley, G. H. (1975). Analysis of frequency of tests and varying feedback delays in college mathematics achievement. *College Student Journal, 9*, 32–36.
Turney, A. H. (1931). The effect of frequent short objective tests upon the achievement of college students in educational psychology. *School and Society, 33*, 760–762.
Ward, E. F. (1984). Statistics mastery: A novel approach. *Teaching of Psychology, 11*, 223–225.
Wiggins, J. A. (1968). *Learning contingencies in the college classrooms: A pilot study.* Final report. (ERIC Document Reproduction Service No. ED 024 314)
Wilkins, S. A. (1979). A study of the effects of various teaching-testing frequencies on cognitive gains. *Dissertation Abstracts International, 40*, 1977A. (University Microfilms No. 79 21993)
Williams, R. L., & Lawrence, J. (1974). *The effects of frequency of quizzing in a lecture course.* Paper presented at the Second National Conference on Research and Technology in Higher Education, Atlanta.

U.S. Postal Service
STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
Required by 39 U.S.C. 3685)

1A. Title of Publication	1B. PUBLICATION NO.	2. Date of Filing
JOURNAL OF EDUCATIONAL RESEARCH ISSN #0022-0671	2 7 9 7 6 0	9/30/91
3. Frequency of Issue	3A. No. of Issues Published Annually	3B. Annual Subscription Price
Bimonthly	6	Institutions: $62. Individuals: $31

4. Complete Mailing Address of Known Office of Publication *(Street, City, County, State and ZIP+4 Code) (Not printers)*
1319 Eighteenth Street, NW, Washington, DC 20036-1802

5. Complete Mailing Address of the Headquarters of General Business Offices of the Publisher *(Not printer)*
1319 Eighteenth Street, NW, Washington, DC 20036-1802

6. Full Names and Complete Mailing Address of Publisher, Editor, and Managing Editor *(This item MUST NOT be blank)*
Publisher *(Name and Complete Mailing Address)*
Walter E. Beach, Helen Dwight Reid Educational Foundation
1319 Eighteenth Street, NW, Washington, DC 20036-1802
Editor *(Name and Complete Mailing Address)*
Board of Executive Editors
1319 Eighteenth Street, NW, Washington, DC 20036-1802
Managing Editor *(Name and Complete Mailing Address)*
Jeanne Bebo
1319 Eighteenth Street, NW, Washington, DC 20036-1802

7. Owner *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated.) (Item must be completed.)*

Full Name	Complete Mailing Address
Helen Dwight Reid Educational Foundation	1319 18th St., NW, Washington, DC 20036-1802

8. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages or Other Securities *(If there are none, so state)*

Full Name	Complete Mailing Address
NONE	

9. For Completion by Nonprofit Organizations Authorized To Mail at Special Rates *(DMM Section 421.12 only)*
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes *(Check one)*
(1) [X] Has Not Changed During Preceding 12 Months (2) [ ] Has Changed During Preceding 12 Months *(If changed, publisher must submit explanation of change with this statement.)*

10. Extent and Nature of Circulation *(See instructions on reverse side)*	Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
A. Total No. Copies *(Net Press Run)*	3,758	3,700
B. Paid and/or Requested Circulation 1. Sales through dealers and carriers, street vendors and counter sales	-	-
2. Mail Subscription *(Paid and/or requested)*	3,283	3,289
C. Total Paid and/or Requested Circulation *(Sum of 10B1 and 10B2)*	3,283	3,289
D. Free Distribution by Mail, Carrier or Other Means Samples, Complimentary, and Other Free Copies	75	75
E. Total Distribution *(Sum of C and D)*	3,358	3,364
F. Copies Not Distributed 1. Office use, left over, unaccounted, spoiled after printing	400	336
2. Return from News Agents	0	0
G. TOTAL *(Sum of E, F1 and 2—should equal net press run shown in A)*	3,758	3,758

11. I certify that the statements made by me above are correct and complete
Signature and Title of Editor, Publisher, Business Manager, or Owner
Gordon Massey Assistant Director