International Journal of Human Sciences
ISSN:1303-5134

Volume: 6 Issue: 2 Year: 2009

Effectiveness of frequent testing over achievement: A meta analysis study

Gülşah Başol*
George Johanson**

Abstract

In current study, through a meta-analysis of 78 studies, it is aimed to determine the overall effect size for testing at different frequency levels and to find out other study characteristics, related to the effectiveness of frequent testing. 78 studies met the inclusion criteria out of 118 experimental and quasi-experimental studies in the potential study pool and 37 of them were journal articles, 21 doctoral dissertations, 14 master theses, and the rest were ERIC documents and conference papers. The newest study in the pool dated back to the year 2003. After the coding process, Hedges' d effect size was calculated. The results were analyzed by SPSS and Meta Win. 233 effect sizes were calculated from 78 studies. Studies with similar exam frequency were categorized into three categories: low, medium, and high according to the frequency of the tests used in the study. According to the results, the cumulative mean effect size for 78 studies was .46. The results also indicated that there were not statistically significant differences between the mean effect sizes of the studies examine students at high, medium and low frequency levels. Overall, the findings indicated that frequent testing increases academic achievement. Also, the effectiveness of a set of categorical variables possibly related to the overall effect size for academic achievement is sought through Q statistic.

Key Words: Meta Analysis, Frequent Testing, Exam Frequency.

*Gaziosmanpaşa Üniversity, basol@gop.edu.tr
**Ohio Üniversity, johanson@ohio.edu

Introduction

Educators are aware of the need for better instructional methods to increase student achievement and are constantly exploring new methods to help students succeed (Palmer, 1974). Repetitive evaluation of subject matter, called "frequent testing", is also one of the ways to improve student's learning experiences. Frequent testing refers to testing within shorter periods than the commonly used two or three midterms and final exam type evaluations. It is important to achieve a proper balance between teaching and testing in order to get both jobs done, effectively. Although the time between the tests varies among frequent testing studies, they all focus on improving student achievement through frequent evaluations and keeping students on the ball.

There have been many studies conducted on the effectiveness of frequent testing since 1920s. The results of these studies often supported the idea of improved outcomes in student performance as the testing frequency increases (Kulik, Kulik, & Bangert-Drowns, 1986). Starting in the 1920s, the effectiveness of frequent testing became an interesting question and many studies followed each other. (Bess, 1938; Curo, 1963; Deputy, 1929; Eurich, Longstaff & Wilder, 1937; Fitch, Drucker, & Norton, 1951; Jones, 1923; Schutte, 1925; Turney, 1931; Williams, 1931). Although the subject matter, school, and grade levels selected as target populations may vary in these studies, the majority of the studies claim that students who are tested periodically make somewhat higher scores on final examinations than students who are not tested periodically (Stanlee & Popham, 1960). Majority of frequent testing studies were carried on U.S college-level students. However, there are also a few studies conducted abroad with similar methodology to the ones conducted in the United States. It is possible to see a variety of testing frequencies from study to study throughout the literature. In addition to differences in frequencies, there were also methodological variations among studies. For example, in some studies frequent testing were used as a part of a learning model while in some other studies it was used solely to see its effect on achievement, not in the concept of any other model. Frequent testing is a main feature of both Bloom's mastery learning model and Keller's PSI and these two well-known mastery-learning models commonly employ frequent testing (Kulik, Kulik, & Bangert-Drowns, 1990).

Regardless of how frequently they are used, tests serve as an important function by examining students' responses to specific questions or problems. Most of the studies considered tests either as formative or summative according to their purpose. Whether tests were used for formative or summative purposes, they all hypothesize that frequent testing had a positive effect upon academic achievement. While some studies focused on the effect of frequent testing upon achievement, others also explored the possible effects over anxiety, attitude and retention.

Throughout the literature examined, frequent testing was reported as a factor increasing academic achievement. However, there were also a few studies concluding that frequent testing had no effect or only a slight effect upon students' academic performance. These conflicting results create confusion and leave questions about the use of frequent testing, whether in formative conditions or summative conditions, unanswered.

As mentioned previously, the effects of frequent testing have been studied for almost a century and there is a massive body of research in need of a comprehensive review. No matter if it is narrative or statistical, literature reviews give opportunities to explore the subject matter more thoroughly and comprehensively. There are a couple of narrative reviews on frequent testing that identify methodological problems or unexpected results throughout the studies (Peckham & Roe, 1977; Gocmen, 1997). There is also a meta-analysis study on the effectiveness of frequent testing upon achievement and it was conducted by Bangert-Drowns, Kulik & Kulik (1991).

According to the results of Bangert-Drowns et al. (1991)'s meta-analysis study, most of the summative studies on the effect of frequent testing report a positive effect of frequent testing upon achievement. While almost all of the studies report a positive effect on achievement, the combination of these results under Bangert-Drowns et al. (1991) meta-analysis fail to indicate a considerable effect size of frequent testing upon achievement.

The Purpose of the Study

This study attempts to synthesize all studies of frequent testing over achievement through a comprehensive, well designed, and up-to-date meta-analysis. The main purpose of the study was to determine an overall effect size for frequent testing on achievement in secondary schools and at college level in the United States. Second, it was determined whether the effectiveness of frequent testing was related to some selected study characteristics such as course subject, school level, presence of feedback, and instructor effects.

As another important point, the most current study in Bangert-Drowns et al. (1991) meta-analysis was dated back to 1989. Since then many studies were done on the topic and still there were many conflicts preventing researchers from believing one way or the other on the effect of frequent testing upon achievement. In addition, database searches indicated that these studies did not cover all existing literature at the time they were conducted. All these suggest a need for a new meta-analysis to include studies that were more recent and to provide a more comprehensive look at the effectiveness of frequent testing on academic achievement.

It is important to see what has changed since the last meta-analysis study in 1991. The current study is a more comprehensive and up-to-date study conducted on the effectiveness of frequent testing with the newest study in the potential study pool dating back to 2003.

Research Questions

The main question of the study is: What is the overall effect of frequent testing on achievement?

Also other questions are formed regarding the substantive study characteristics and study features in order to find out whether there were effect size differences according to this variables. The study charahteristics were whether the frequent testing used in the frame of mastery learning or not, frequency of testing in the experimental group, instructor effect, sample size, amount of feedback received between experimental and control group, the use of standardized tests or teacher made tests, the use of objectively scored tests or not, the use of factual items or items requiring higher order thinking skills, and formative or summative use of test. Study features were duration of the tratment, assignment of the subjects to the groups, school level, and subject matter.

There were 13 questions to be answered regarding the effect of frequent testing according to the some study characteristics and features. The questions regarding the substantive characteristics and study features were formed as the following:

1. Does the effectiveness of frequent testing differ when it is used in the frame of a mastery learning model or when it is used alone as a teaching aid?
2. Does the effectiveness of frequent testing differ by the frequency of tests in the experimental groups? And so forth.

Three more questions were added to be cautious against publication bias and time factor on the topic of interest.

1. Does the effect size differ by the year of the report?
2. Does the effect size differ by the publication type?

3. Is there a publication bias in the literature on the subject of the effectiveness of frequent testing on student achievement? In other words, are there effect size differences between published and unpublished (e.g. conference papers, thesis and dissertations) studies?

Methodology

In order to locate all research examining the effect of frequent testing on achievement, several approaches were used for this meta-analysis. The majority of the studies were located through a computer bibliographic search by several electronic databases. These databases were ERIC (Educational Resource Information center), Social Science Citation Index (SSCI), PsycINFO, Education Abstracts, Digital Dissertation Index, PROQUEST, and Worldcat. Beside computer database searches, a search was undertaken to get a hold on the studies that may not have been included in the computer search databases. This search was conducted through the journals in which 80% of the studies on the effectiveness of frequent testing have been published: *Journal of Educational Research, The Psychological Record, Journal of Educational Measurement, The Journal of Educational Psychology, Journal of Experimental Education, Science Education, and School and Society.* Furthermore, a manual search was conducted of the reference lists of each study that was included in this meta-analysis were cross checked for additional articles missed by database searches. Subsequently, educational sites on Internet were used as an attempt to locate relevant research studies.

Procedure

The sample consisted of studies examining the effectiveness of frequent testing on student's academic achievement. The search encompassed studies published between 1929 and 2003 and Unpublished Doctoral dissertations and master's theses completed during that period. One hundred and eighteen studies were found and included in the potential study pool. Once all searches were completed, these studies were checked against certain inclusion criteria. In addition to 35 studies included in Bangert-Drowns et. al. (1989) study, 71 studies were included in the current analysis that were conducted 1989 and before. Thirteen of 118 studies were conducted after 1991. A critical view is employed for inclusion and exclusion of the studies, which is a requirement for a well-done meta-analysis (Glass, McGaw & Smith, 1981).

Inclusion Criteria: After the pool of potential articles was formed, all candidate studies are screened against the following criteria.

- The experimental and quasi-experimental studies on the effect of frequent testing on achievement, conducted in the United States on secondary education and college level students,
- The studies with sufficient data (means, standard deviations, number of subjects) for effect size (d) calculations, and
- The studies not reporting an effect size, but some parametric statistics such as "*t*" and "F" test results, means and standard deviations reported were included.

Exclusion Criteria: The exclusion criteria were as follows:

- The studies with only qualitative findings were excluded from the current study due to insufficient data to calculate the effect size.

After determining the potential pool of studies to be included, and getting ready for the actual coding through a pilot coding study of five articles by the researchers and two coders, an extensive coding process was carried.

Coding Process

There were different strategies for retrieving predictors of study outcomes. These were direct coding, judge's rating the quality of the studies, by using post-hoc theoretical indices, and predictors derived from archival and historical sources (Mullen, 1989). In the current meta-analysis, predictors were derived through direct coding and some additional predictors that were also used in the previous meta-analysis studies on the related topics. By using a coding form, all critical study information was translated into a coding form. Through these forms, the information related to the methodology of the study and relevant to the topic of the study was collected. The characteristics regarding procedures, experimental design, settings of the study, publication histories can be diagnosed through careful coding of studies.

Reliability of Coding

It is recognized that error occurs frequently in coding processes. In order to obtain a certain level of reliability, a pilot of five studies were coded to check the inter-rater reliability between the raters and also to make the necessary corrections to the coding form. Two Ph.D. students and the researcher at college of education of Ohio University served as the possible coders. After coders independently coded a subset of five articles from the study pool, Cohen’s кappa statistics was used to determine the reliability among coders. Cohen’s кappa is a measure of reliability corrected for chance occurrence (Landis & Koch, 1977). Values of кappa above .75 are considered to be representative of an excellent agreement among coders, while 0.40-.75 is to be considered to be fair to good agreement beyond chance (Landis & Koch, 1977). An overall κ of .89 was obtained for the two coders and the five studies for this pilot. In other words, there is an 89% percent agreement among raters across items beyond chance agreement. The result indicated that the reliability among coders was very high. Final revisions were made to the coding form according to the feedback from the pilot study.

As an attempt to increase the reliability of the coding, the researcher coded each five set of the studies before the coders. After coders returned the coding forms, the results of their coding were compared to the results of the researcher’s coding. If there was a disagreement, the researcher and coder discussed and solved any problems in the understanding before further coding of studies. After coding forms were entered to data file, the researcher checked each study’s coding to make sure that no mistakes was made during the data entering process. This way, it was expected to increase reliability of the coding process. After the discussions, it was decided that if there was still continuous disagreement between the researcher and the coders, data for that particular item was decided to be counted as missing.

Disagreements on the sample studies were discussed and necessary changes were made to the coding form in order to eliminate further problems in the coding process. Coding through a reliable set of rules plays an important role to establish the reliability of an analysis. Glass et al., (1981) emphasize the importance of reliability in a meta-analysis and emphasize it as the biggest problem of meta-analysis.

Validity of the Study

Validity refers to the agreement between the value of a measurement and its true value. Validity is quantified by comparing measurements with values that are as close to the true values as possible. Poor validity also degrades the precision of measurement, and it reduces the researcher's ability to characterize relationships between variables in descriptive studies.

Thoroughness of literature search; selection of studies for inclusion; appropriateness of coding and analysis of studies are some of the factors affecting validity in a meta-analysis. It is reported that mixing the outcomes from the rigorous and non-rigorous research to obtain a common result may result in poor validity in a meta-analysis study (Eysenck, 1978; Glass, et al., 1981). Cooper (1998)'s suggestions in order to obtain validity was applied in the current meta analysis. First of all, the same criteria must be used for inclusion and exclusion of the studies regardless of their findings on any other characteristics. For example, if there was not any information regarding the effect of frequent testing on achievement, the studies, which provided information to calculate the effect sizes for aptitude, anxiety, and retention, were excluded from the analysis. In addition, the same exclusion criteria were performed on the calculation of effect sizes for every single study question related to dependent variables and effect sizes. To be able to define studies as well as possible, all representative characteristics were listed in the coding form and necessary definitions were given in the variable list. Light (1980) also expresses the importance of reliability of judgments in the selection and acceptance of sources. Since reliability is necessary for validity, it is important to deliver study characteristics to the coding form consistently and accurately. Without reliability, there would not be any concern for validity. As another precaution for validity, the unit of analysis is defined as the effect sizes calculated for each outcome variable from the pool of studies. If a study was conducted with the same sample at different times, only one of these was included. In this way, overlooking or exaggerating the same study's effect was prevented.

Variables: As in any experimental and correlational studies, there are dependent and independent variables in a meta-analysis. Meta-analysis simply aims to report an overall answer about the effect of an independent variable on a dependent variable. Effect sizes are dependent variables of meta-analysis studies.

Dependent Variables: The dependent variable was student achievement for this meta-analysis. Achievement is defined as an outcome measure for some type of performance (standardized and teacher-made tests, grades, quality of performances such as compositions and presentations, quality of products such as reports, and so forth). A variety of experimental settings and tasks were used in the studies yielding effect sizes for the dependent variable of achievement. The researcher expected to find out if there were additional moderator variables affecting the dependent variable "achievement". Throughout the literature, there were studies measuring attitude, anxiety, and retention of the material, and study time as additional outcome variables. These variables were not included in the study because of the excessive number of the studies and time and money constrain.

Independent Variables: In a meta-analysis, the independent variables are study descriptors. Coding forms were used to identify these study characteristics. The factors, that were coded in the coding form was reported statistics for each study (standard deviations, means, effect sizes, *t* tests, *F* tests, correlations, chi squares, degrees of freedom), sample size and variables related to the substantive and methodological characteristics such as frequent testing in the frame of mastery learning or not, frequency of tests for experimental group (high, medium or low frequency), duration of treatment (as the number of weeks study lasted), subject assignment (whether subjects are randomly assigned to groups or not), instructor effects (studies using the same instructor to teach both experimental and control group or not), feedback (feedback is present in experimental group, in control group or not present in either of the groups), nature of assessment instrument (teacher made, instructor made, or standardized tests), objectivity (selected response, objectively scored tests or constructed response subjectively scored tests), level of skills required in the test items (factual items or items requiring higher order skills, or mixed), instructional role of the test (formative or summative), school level (college or secondary education), and subject matter (education, psychology, mathematics, physics, chemistry).

Characteristics related to publication histories of the study were publication year (database search for the year 1920 and 2009 were yielded 118 studies between 1929-2003), publication type, publication source (Journal article, ERIC document, dissertation, conference paper, and so on).

Statistical Analysis

Meta-analysis is a statistical reviewing method to re-evaluate the findings of the studies in order to provide feedback for future research. In order to conduct a meta-analysis on a topic, effect sizes need to be calculated for each study and each experimental group. Effect size, denoted by symbol "*d*", is a standard mean difference between the experimental and control groups divided by a pooled standard deviation. Cohen (1988) defined effect sizes as "small, *d* = .2," "medium, *d* = .5," and "large, *d* = .8", stating that "there is a certain risk in inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (p. 25).

Effect size is a way of quantifying the difference between the experimental and control groups. In this study, effect size is a measure of the effectiveness of frequent testing. There are several ways to calculate the effect size. The basic formula of Glass's approach is for his "*g*" statistics:

$$g = \frac{\text{Mean of experimental group - Mean of control group}}{\text{Standard deviation of control group}}$$

Because Glass's g formula is found biased with small sample sizes; Hedges and Olkin (1985) suggested another effect size estimation formula with pooled within-group standard deviation instead of standard deviation of control group in the denominator. In the current study, Hedges'*d* formula was used for the estimation of effect sizes. (Hedges & Olkin, 1985). Hedges'*d* is derived from Hedges'*g*, also known as Hunter and Schmidt's *d*, with a simple correction against the effect of small sample sizes on the estimated effect size measure.

$$g = \frac{\overline{X}_e - \overline{X}_c}{S_p}$$

Hedges & Olkin, (1985).

The corresponding formula for the pooled standard deviation.

$$S_{pooled} = \sqrt{\frac{(N_e - 1)\, S_e^2 + (N_c - 1)\, S_c^2}{(N_e - 1) + (N_c - 1)}}$$

Hedges & Olkin, (1985).

In some studies the standard deviation and the means of the separate groups were not reported, if this is the case, effect size was calculated for reported values of *t*, *F*, or for *r* statistics. In order to obtain a uniform effect size over the studies, after the conversions of *t*, *F*, and *r* values, Cohen's *d* values first converted to Hedges'*g* then reconverted to Hedges'*d*. The detailed information about the conversion formulas can be obtained from Başol-Göçmen (2004). So far, the data were converted to a common statistic, which was Hedges'*d*. The random effect model, strongly advocated by Hunter and Schmidt, was used in the current meta-analysis. As it is required in a meta analysis, the variance for each effect size was calculated (Rosenberg, Adams, & Gurevitch, 2000). Effect sizes and their variances were calculated by using SPSS for each comparison group in the studies. In the sample, there were studies with multiple comparison groups. Since the effect sizes, coming from different groups of the same study, are dependent, multiple effect sizes would cause the over-representation of the same study and this could be misleading, therefore individual studies were decided to be the unit of the analysis (Lipsey & Wilson, 2001). The final data set included the mean effect sizes per study and their variances. Aggregation feature in SPSS was used to create the mean effect size per study. The further analysis was carried in Meta Win 2.0, a software program specifically designed to perform meta-analysis. A weighted average effect size was calculated to estimate a cumulative effect size, then the cumulative effect size ($\overline{\overline{E}}$) , and the variance of the cumulative effect size ($s^2_{\overline{\overline{E}}}$), and the confidence interval around $\overline{\overline{E}}$ (refer to Rosenberg, Adams, & Gurevitch, (2000) for the formulas). While calculating the confidence interval for effect size estimation, a *t*-statistic was used, because of its appropriateness for small sample sizes, which is often the case in meta-analysis.

Testing for Homogeneity of Effect Sizes

Before pooling the estimates of effect size from a series of *k* studies, it is important to determine whether the studies can reasonably be described as sharing a common effect size (Hedges & Olkin, 1985, p.122). The following is the hypothesis for the homogeneity of effect sizes.

$H_0 = \delta_1 = \delta_2 = \ldots = \delta_k$

The null hypothesis, δ_i the population Hedges' d effect size, is tested against the alternative hypothesis that at least one of the effect sizes δ_i will be different than the rest (Hedges & Olkin, 1985).

In order to detect the total heterogeneity of a sample, a *Q* statistic was used. Meta Win 2.0 calculates the Q_{Total} , total heterogeneity, and also Q_{wj}, heterogeneity within each group. The *Q* statistic is distributed as a chi-square distribution with *k*-1 degrees of freedom where *k* is the number of effect sizes (Hedges & Olkin, 1985). If the *Q* test is significant, the null hypothesis of homogeneity must be rejected and this means that the variability across the effect sizes is greater than is expected from subject-level sampling error alone (Lipsey & Wilson, 2000). Therefore, each effect size does not estimate a common population mean. Lipsey and Wilson suggest the use of a random-effects model in the case of heterogeneous effect sizes assuming that the variability beyond subject-level is random in one condition, if the sample sizes are not small. In the present study, Hedges and Olkin's *Q* statistic was used to test for the homogeneity of the studies because of availability and because of its certainty.

The value of total homogeneity can be calculated through the summary analysis in Meta Win 2.0 and the analysis is listed under the title "heterogeneity". If the null hypothesis of homogeneity of the effect sizes across studies is rejected, this means that the effect sizes are not homogeneous.

Results

This study's main purpose was to determine the overall effect size for testing at different frequency levels and also to find out if some other study characteristics were related to the effectiveness of frequent testing through a meta-analysis study.

After the data-entering process was completed, Hedges' *d* effect size was calculated for the studies with reported mean and standard deviation information for the experimental and control groups. Hedges' *d* formula requires the calculation of the pooled variance and the group mean difference between experimental and control groups (equivalent to the pre-test group in studies with pre-test- post-test design where there is no treatment or control group mentioned) and it is also corrected for sample size bias. For the studies providing *F, t,* and *r* test values rather than group means and standard deviations, Cohen's *d* statistic was calculated through a set of conversion formulas, then these Cohen's *d* values were converted to Hedges' *g*, and finally they were converted to Hedges' *d* statistics. Later, the variances of each effect sizes were calculated. Finally, the mean effect sizes and the variances for each study were calculated through the aggregation function in SPSS.

After the aggregation, the data set contained one effect size value and its variance for each study in the sample. The rest of the analyses were carried in Meta Win 2.0, statistical software. Through the use of Meta Win, the summary analyses were performed in order to have an understanding of the overall cumulative mean effect size for the studies, their variances and their 95% confidence interval. The homogeneity of mean effect sizes hypothesis was tested by the use of *Q* statistic.

Descriptive Statistics

The initial literature review yielded 118 studies that were possible candidates of study sample of the following analyses. Of 118 studies, 78 were retained in the final study sample. The following table provided the number of studies in the initial and final study pool from each publication group.

Table 1
The Number of Studies in the Previous Study Pool Versus the Remaining Studies in the Final Study Pool According to Publication Type

Publication Type	Number of studies before coding	Number of studies included after coding	Number of studies excluded	Percentage of included studies
Journal articles	55	37	17	67.27 %
Dissertations	29	21	8	72%
Seminar papers	2	1	1	50%
Master’s thesis	21	14	7	66.66%
Technical reports	4	1	3	25%
ERIC documents	3	2	1	66.66%
Conferences papers	4	2	2	50%
Total	118	78	39	66.10%

Of 118 studies, 78 studies were kept and 40 studies were not included in the final data set. There were several reasons for withdrawing these studies. The main reason for not including some of the journal articles was that these studies did not meet the inclusion criteria. According to the inclusion criteria, studies with insufficient data to calculate the effect size, studies with a dependent variable other than achievement were to be excluded from the final data set.

The initial study pool consisted of 29 dissertations and 21 master’s thesis and these studies were identified as possible sources of information in the final study pool. However, 8 of the dissertations and 7 of the master’s theses were not included in the final study sample. These dissertations and master’s thesis were not included mainly because of the lending restrictions of the institution that owned the study. Ten out of 40 disqualified studies were not included because they were not related to achievement, which was the dependent measure in the study. Another ten studies from the initial study pool were withdrawn because of the lack of required information to calculate an effect size. There were a few studies that were not included in the final sample for other reasons, such as, duplicate studies (more than one publication form), studies at a school level other than college, studies conducted out of the United States, and so on.

The studies in the sample of this research were coming from variety of disciplines. The number of effect sizes per study also varied. As mentioned before, 78 studies were included in the final study sample. From these 78 studies, a total of 233 effect sizes were calculated. A list of the studies in the initial study pool is given in Appendix A. Appendix B contains the coding form and Appendix C contains the variable list. Appendix D provides subject area information for each study and the number of effect sizes, calculated from each study. Appendix E is intended to provide descriptive information about the studies in the final study pool and their subject areas and the number of effect sizes calculated from each study. Appendix E also provides descriptive information regarding the independent variables in the study and the percentages of effect sizes within each category.

Studies with similar exam frequency were categorized into three different categories: low, medium, and high according to the frequency of the tests used in the study (other than the final exam). Studies that used daily, every other day or less than weekly exams were coded as high frequency; studies with weekly exams were coded as medium frequency; and the studies with every other week and less frequent exams were coded as low frequency. The frequency and the percentages of studies in the sample of this meta-analysis study are given in Table 2.

Table 2
The Percentage of Studies in Each Frequency Group for the Data with Mean Effect Size Per Study and the Data with an Effect Size for Per Comparison Group

		Frequency	Percent	Frequency	Percent
Valid	High	22	28.2	64	27.5
	Medium	41	52.6	125	53.6
	Low	15	19.2	44	18.9
	Total	78	100.0	233	100

According to Table 2, of the 78 studies, 22 used daily or every other day exams, 41 used weekly exams, and 15 used every other week or less frequent exams in the experimental group. Sixty-four effect sizes out of 233 were coming from high frequency studies, 125 were from medium frequency level, and 44 studies were coming from low frequency group. The majority of the studies in the sample used weekly tests in their experimental groups.

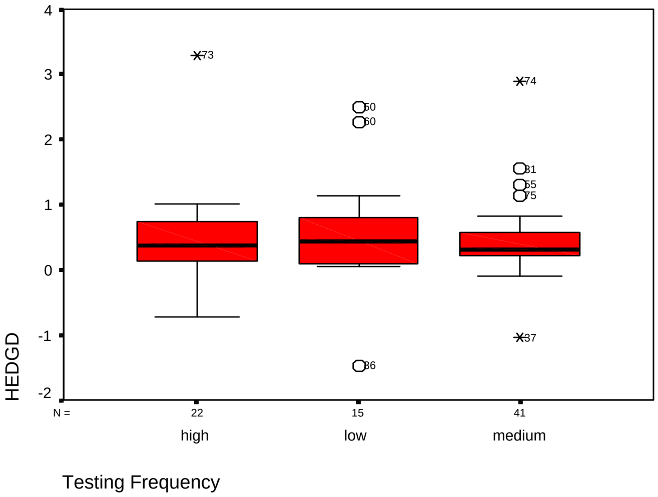
Search for the Outlier Studies

Treating multiple effect sizes from the same study as if they were coming from separate studies can misrepresent the real effects of frequent testing. A simple count revealed that almost 190 effect sizes out of the 233 were equal or greater than .5, while 30 effect sizes out of 78 studies were equal or greater than .5. An overall analysis was conducted with both data sets for reasons of comparison. The data set with the 233 effect sizes was not used for further analysis because the over representation of the sample could be threat to the independence of observations assumption. After the overall analysis, the remaining analyses were carried out using the data set with Hedges' weighted mean effect size.

Through the visual inspection of the stem and leaf plots, a funnel plot, and normal quantile plot, it appeared that Townsend study (1972) was an outlier with its especially large effect size of 3.30. Townsend study was removed from the data to see its impact. After removing the study, four analyses (duration of treatment, feedback, assessment type of final exam, and study's subject area) turned out not significant while they were significant before. Only one analysis (nature of final assessment) changed from not significant to significant. Therefore, the further analyses run without this study.

A box plot of effect sizes for frequency level appears in Figure 1.

Figure 1
Box Plot of Effect Sizes by Testing Frequency



According to this box plot, studies with medium and low exam frequency had a few more extreme cases than the high frequency group.

Overall Effectiveness of Frequent Testing on Academic Achievement

The cumulative mean effect size for 232 effect sizes out of 77 studies, was .41 as it was derived from weighted Hedges' *d* effect size calculations. This means that the frequent testing raised academic achievement scores by .41 standard deviations. In other words, assuming a normal distribution of scores an average student receiving frequent tests did better than 66% of the students that are not exposed to frequent testing. Ninety-five percent confidence levels were found to range from .32 to .49. There was one study with a 0 effect size value indicating no difference on the academic achievement between the weekly tested and monthly or less frequently tested group. Furthermore, among the studies with negative effect sizes, there was a study with a large negative effect size of –2.5.

The cumulative mean effect size for 77 studies was .44. According to the mean effect size for each study, the frequent testing raises academic achievement scores by .44 standard deviations. Ninety-five percent confidence level ranged from .29 to .58. In this group of 77 effect sizes, there were 9 studies with a negative effect size and the remaining 68 studies had a positive effect size.

Homogeneity of the Effect Sizes

In order to analyze the variance in effect sizes across the studies, Q statistic was used to test the homogeneity of the effect sizes. In addition to the homogeneity analyses, Hedge's mean effect size values for each level of every variable and their confidence intervals were also provided. As can be seen from the corresponding tables, some of the confidence intervals included zero, which may indicate a true null hypothesis.

Effect Sizes by Frequency of Testing

The hypothesis examined was whether the frequency of testing have an impact on the academic achievement. The frequency of testing variable had three levels: high frequency group, medium frequency group, and low frequency group.

Table 3

Effect Sizes by Frequency of Testing (N=77)

Variable: Frequency of Testing

Heterogeneity tests	*p*	*df*
Q_{Total} = *108.0693*	*0.00919*	76
Q_{Bet} = *0.8378*	*0.65777*	2
Q_{Wi} = *107.2315*	*0.00699*	74

According to Table 3, Q_{total} value of *108.07* was significant (p = *0.00919*). Therefore, the null hypothesis of homogeneous effect sizes was rejected. This suggested that the variability of Hedges' *d* weighted mean effect size was different from that expected by sampling error. Between class effects were also found to be heterogeneous, which indicated a significant amount of variance within the groups remained unexplained (Q_{Bet} = *0.838*, p = *0.657*). In other words, there were not statistically significant differences between the mean effect sizes from three study groups, using tests at different frequency levels. Each group had similar mean effect sizes.

When the homogeneity of the effect sizes in each level was reviewed, Hedge's mean effect sizes for high and medium frequency group were found to be homogenous with a non-significance Q_{Total} = 23.34, df = 20, p = 0.27 for high frequency group, Q_{Total} = 51.21, df = 40, p = 0.11 for medium frequency group). The mean effect sizes of the studies in the low frequency group found to be heterogeneous with a significant Q_{Total} value of 27.54 in this group of 15 studies (p = 0.016). The mean effect sizes and their 95% confidence intervals for each frequency level were given in the Table 4. According to the results, the mean Hedge's *d* effect sizes were very similar in each group. The studies in medium frequency group had a smaller effect size confidence interval despite the larger number of studies in this level compared to the studies in high and low frequency group.

Table 4
Mean Effect Size and 95% CI in Each Group (N=77)

Class	df	Mean d	95% CI for d
High	20	0.3596	0.1786 to 0.5407
Medium	40	0.4215	0.2706 to 0.5723
Low	14	0.5227	0.1753 to 0.8700

According to the results in Table 4, the cumulative mean effect size in the high frequency group was 0.3596 and the lower bound for 95% confidence interval was 0.179 and 0.541 for the upper bound, cumulative mean effect size for the medium frequency group was 0.422 with 95% confidence interval from .271 to .572, and finally, cumulative mean effect size for the low frequency group was .523 and the confidence interval around this mean ranged from 0.175 to 0.870. Of the 77 studies, 41 studies used medium frequency tests, which were weekly, 21 studies used less frequent tests than weekly tests and 15 studies tested their students in the experimental group more frequently than weekly.

Results

When the findings from this research are compared to the ones from Bangert-Drowns, Kulik and Kulik (1991) study, the only prior meta-analysis study on the topic, it is found that the current study resulted in a larger mean effect size. The Bangert-Drown et al.'s meta-analysis of 35 studies reported a mean effect size of 0.23. In Bangert-Drowns et al.' study, 29 studies (83%) reported positive results, while 6 studies (17%) reported negative effect sizes. The current study resulted in 69 positive mean effect size (89%) value and 9 negative ones (11%) out of 78 studies. Although not quite the same, overall the results of this study supported the findings of Bangert-Drowns et al. (1991) meta-analysis.

The findings of the current study failed to explain the source of variation among the data. The 78 studies seemed to be heterogeneous in their mean effect sizes for almost all moderator variables, indicating overall unexplained differences among them. However, the between effect variance was found to be very low and non- significant, while within groups variation was quite high and significant. This indicates that the moderator variables were not able to explain the difference among means. The results for each of the analyses follow:

The first moderator variable "mastery learning" was intended to search if the effectiveness of frequent testing varies among the studies using frequent testing as a part of a mastery learning model, Keller's PSI or Bloom's Mastery learning mainly, and also the studies using frequent testing as a teaching methodology without in the context of any other learning model. Overall, studies in the different categories are found to be heterogeneous, the question of whether there is a difference on the effectiveness of frequent testing according to its use as a part of a mastery learning model or not remained unanswered because of a non-significant between groups heterogeneity value.

The second moderator variable "duration of treatment" is found to be as one of the rare factors affecting the effectiveness of frequent testing though there was much larger unexplained variation among the studies. Surprisingly, the studies lasted less than a quarter had the highest mean effect size value. One might question whether this is because of not having enough time to see the real effectiveness of frequency or as the result of the Howthorne effect, everlasting excitement of a new application. Much to the surprise, the studies lasted longer than a semester had considerably small effect size. It is hard to detect of the impacts of this short-term effect after some time. This explains the reason behind having a low mean effect size value for the studies lasted longer than a semester. Besides, one might argue how efficient it can be to switch the testing frequency within a quarter considering how difficult time management is within the quarter system.

The variable "subject assignment method" were not found significant, either. According to the findings, effect sizes did not vary significantly by subject assignment method. One interesting finding from the analysis is that although it was not significant the studies that the experimental and control group were formed by a random assignment with a pre-test had almost twice as large mean effect size value compared to the studies that experimental and control groups were formed without the use of a pre-test. However, in contrary the mean effect size for the studies that failed to use a random assignment method were larger than both.

Interestingly, the variable instructor effect and the instructional time effect were both found to be non-significant on their variability in the effect sizes. Unlike the expectations, the studies that used the same instructor to teach both experimental and control group had a smaller mean effect size value compared to the studies using different instructors. However, it is logical to claim that the studies using the same instructor in their design to teach both experimental and control groups are methodologically sounds better.

The effect size did not vary significantly in the variable “sample size” either. Interestingly the studies with smaller sample sizes had the largest mean effect size value. Much to surprise, the studies with 100 or more subjects in their sample had the second largest mean effect size value, which does not make any sense considering the highest mean effect size value belonged to the smallest sample size.

Variable “feedback” is found to be significant which means that effect sizes varied significantly by this variable. When the mean effect sizes are reviewed, the group of studies, providing feedback to their experimental group are found to have a larger mean effect size value, which suggesting a bias on the methodological design of these studies in the favor of experimental group.

The effect sizes on the variable “nature of the assessment” did not seem to vary significantly. Not much to surprise, the studies using teacher-developed tests had the largest mean effect size, which suggesting a teaching to test effect since the teachers were the ones developing the test material. As it is expected, the studies using standard tests had a large effect size, too.

Furthermore, it is found that the categories within the variable “assessment type” significantly vary in their effect sizes. In this group the studies using the multiple-choice tests had the largest mean effect size value, although they were little in their quantity. However, the majority of studies used objective tests and according to the Cohen’s criteria, their mean effect size can be considered medium.

The effect sizes differences in the levels of the variable “conceptual level” were not significant, either. Among the levels of this variable (factual, conceptual, problem solving), the studies using conceptual level items had the largest mean effect size value. However, more than half of the studies in the sample did not report the conceptual level of the items that were used.

The instructional level of the test was one of the main inquiries of the study. It would be interesting to see if the effect sizes would vary according to grading or not grading the frequent classroom assessments. The findings suggested that there was not any differences between the studies using frequent graded tests or studies using frequent not graded tests. Although frequent testing is a big part of the major mastery learning models and they were encouraged to be used to monitor students' weaknesses and strengths, the practice seemed to be different than the ideology considering that the majority of the studies used graded frequent tests. These findings contradicts with the expectation of higher grades from the studies of frequent summative testing compared to the frequent formative testing (Stanlee and Popham, 1960).

Although the effect size of the studies did not vary according to their "school level", the majority of the studies were conducted at the college level.

The studies were found to differ in their effect sizes according to the variable "subject matter". Among the levels of this variable, the subject level Math had the largest mean effect size value.

The variable "ability level" was included to see whether using a control variable while forming the experimental and control groups have an impact on the results. The results were not significant and the number of the studies in both levels (studies using statistical control over the results or not) and their effect sizes were similar in magnitude.

The variable "publication year" was not significant, either. Which means the effect sizes of the studies did not vary according to their publication year.

The present meta-analysis study detected significant differences between the levels of "publication type" and the variable "publication status", suggesting that the effect sizes of the studies vary according to their publication type and their publication status. The majority of the studies were journal articles and they had a large mean effect size value. As expected the published studies had a larger mean affect size value. This finding is supported by majority of the meta-analysis studies, suggesting that published studies are resulted in more positive results.

Bangert-Drowns et al. (1991) reported that an association between high frequency testing and low grades. According to the results of the current study, the mean effect sizes among the high, medium, and low frequency groups were found to be similar. This suggests that there is not any difference among the mean effect sizes according to the frequency of testing. In other words, delivering daily tests to the experimental group did not result in a higher mean effect size value compared to using weekly or monthly tests in the experimental group.

Conclusions

The findings of this meta-analysis lead to the several conclusions. First of all, overall frequent testing has a positive effect on academic achievement. Second, the effectiveness do not differ according to the frequency level used in the high, medium and low frequency group studies.

The results of this study indicated that although there were number of significant findings. None of the moderator variables were strongly related to frequency of testing used in these studies. On certain levels, the findings of this meta-analysis were supported by the findings of previous research, such as Bangert-Drown et al.'s meta- analysis study, found a positive effect of frequent testing on academic achievement.

In summary, considering the large magnitude of within effects variation in the search of every research question, though research has uncovered many of the factors that may influence the effectiveness of frequent testing, much remained unanswered. If indeed the effect of frequent testing do not differ among different frequency levels, then it is important for educators to find out the optimal number of non-graded exams to be given in order to increase student achievement. Therefore, the judgment on the effectiveness of frequent testing among the different frequency levels remains to be determined.

Recommendations for Future Study

One big challenge of this meta-analysis study was how to define high, medium and low frequency testing. The same meta analysis study can be conducted in number of ways depend on what is considered as high, medium, and low frequency of testing.

Another challenge was to decide how to incorporate the frequency level used in the control group into the study's independent variables. Since the present study's main interest is to see the impact of the frequency of testing in the experimental group, the studies are not coded according to the frequency level used in their control group. Later on, it is realized that more thought might have given to the frequency of testing in the control group to see whether it has an impact over the results. However, while one frequency level is considered as medium in the experimental group categorization, the same frequency might be used in the control group in another study. Therefore, it is hard to defend that this is a fair comparison.

There were number of studies looking at the effect of frequent testing on number of dependent variables in addition to achievement, such as anxiety, attitude, retention and etc. Future study might also look at the effectiveness of frequent testing on these dependent variables.

In addition to frequency of testing, the influence of certain moderator variables, either related to substantive or methodological study characteristics, on academic achievement needs to be investigated in order to see how these additional moderator variables would explain the relationship between the frequency of testing and the academic achievement.

References

Bangert-Drowns, R. L., Kulik, C-L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213 - 238.

Başol-Göçmen, G. (2004). Meta Analysis. *Eğitim Araştırmaları*, 15, 16-22.

Bess, E. J. (1938). The effects of written examinations on learning and on the retention of learning. *Journal of Experimental Education*, *7*, 55-62.

Cooper, H. M. (1998). Synthesizing research: A guide for literature reviews. Thousand Oaks, CA: Sage.

Curo, D. (1963). *An Investigation of the influence of daily pre-class testing on achievement in high school American history classes*. Purdue University. *Dissertation Abstracts International*, *24*(12), 5236.

Deputy, E. C. (1929). Knowledge of success as motivating influence in college work. *Journal of Educational Research*, *20*(5), 327-334.

Eurich, A. C., Longstaff, H. P., & Wilder, M. (1937). *The effects of weekly tests upon achievement in psychology*. Report of the Committee on Education Research, University of Minnesota. Minneapolis: University of Minnesota Press.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*, 517.

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

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

Gocmen, G. B. (1997). *The effects of frequent testing on achievement*. Unpublished Seminar Paper, Ohio University.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press Inc.

Hunter, J. E., & Schmidt, F. (1990). *Methods of meta-analysis*. Newbury Park, CA: Sage.

Jones, H. E. (1923). Experimental studies of college teaching. *Archives of Psychology, 68*, 5-70.

Kulik, C-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1986). Effects of testing for mastery on student learning. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Kulik, C-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*(2), 265-299.

Landis, J., & Koch, G. (1977). The measurement of observed agreement for categorical data. *Biometrics*, *33*, 159-174.

Light, R. J. (1980). Synthesis methods: Some judgment calls that must be made. *Evaluation in Education*, *4*, 13-17.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Applied social research methods series, 49. London, New Delhi: Sage.

Palmer, E. L. (1974). Frequency of tests and general subject area mastery. *Psychological Reports*, *35*(1), 422.

Pecknam, P. D., & Roe, M. D. (1977). The effects of frequent testing. *Journal of Research and Development in Education*, *10*(3), 40-50.

Rosenberg, M. S., Adams, D. C., & Gurevitch, J. (2000). *Meta Win: Statistical software for meta-analysis*. Sunderland, MA: Sinauer Associates.

Schutte, T. H. (1925). Is there any value in the final examination? *Journal of Educational Research*, *12*, 204-13.

Stanlee, L.S., & Popham, W. J. (1960). Quizzes' contribution to learning. *Journal of Educational Psychology*, *51*(6), 322-325.

Townsend, N. R. (1972). *The relationship of frequency of tests and delay of feedback of test results to achievement in first quarter analytic geometry and calculus*. Purdue University. *Dissertation Abstracts International*, *33*, 06 A.

Townsend, N. R. (1972). *The relationship of frequency of tests and delay of feedback of test results to achievement in first quarter analytic geometry and calculus*. Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.

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.

Williams, M. L. (1931). *The effect of frequent classroom testing on the learning and retention of subject matter*. Unpublished master's thesis, University of Cincinnati.