ED 227 133 TM 830 125

| | |
|---|---|
| AUTHOR | Hedges, Larry V. |
| TITLE | Statistical Methodology in Meta-Analysis. |
| INSTITUTION | ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J. |
| SPONS AGENCY | National Inst. of Education (ED), Washington, DC. |
| REPORT NO | ERIC-TM-83 |
| PUB DATE | Dec 82 |
| CONTRACT | 400-78-000g |
| NOTE | 79p. |
| AVAILABLE FROM | ERIC/TM, Educational Testing Service, Princeton, NJ 08541 ($7.00). |
| PUB TYPE | Information Analyses - ERIC Information Analysis Products (071) -- Guides - Non-Classroom Use (055) |
| EDRS PRICE | MF01/PC04 Plus Postage. |
| DESCRIPTORS | Analysis of Variance; Correlation; Error of Measurement; *Mathematical Models; *Research Methodology; Research Problems; *Statistical Analysis; Statistical Studies |
| IDENTIFIERS | *Effect Size; Glass (G V); *Meta Analysis |

ABSTRACT

Meta-analysis has become an important supplement to traditional methods of research reviewing, although many problems must be addressed by the reviewer who carries out a meta-analysis. These problems include identifying and obtaining appropriate studies, extracting estimates of effect size from the studies, coding or classifying studies, analyzing the data, and reporting the results of the data analysis. Earlier work by Glass, McGaw, and Smith describes methods for dealing with these problems: and has generated a great interest in the development of systematic statistical theory for meta-analysis. This monograph supplements the existing literature on meta-analysis by providing a unified treatment of rigorous statistical methods for meta-analysis. These methods provide a mechanism for responding to criticisms of meta-analysis, such as that meta-analysis may lead to oversimplified conclusions or be influenced by design flaws in the original research studies. Contents include: indices of effect size, statistical analysis of effect size data, assumptions and the statistical model, estimations of effect size, an analogue to the analysis of variance for effect sizes, the effects of measurement error on effect size; statistical analysis when correlations or proportions are the index of effect magnitude, and statistical analysis for correlations as effect magnitude. (Author/PN)

# STATISTICAL METHODOLOGY IN META-ANALYSIS

ERIC/TM REPORT 83

by
Larry V. Hedges

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08541

ERIC/TM Report 83

# STATISTICAL METHODOLOGY IN META-ANALYSIS

by

Larry V. Hedges

The University of Chicago

December 1982

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service, Princeton, NJ 08541

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either these reviewers or the National Institute of Education.

ERIC Clearinghouse on Tests,
Measurement, and Evaluation
Educational Testing Service
Princeton, NJ 08541

## Table of Contents

Introduction....................................................3

Indices of Effect Size..........................................7

The Effect Size as an Index of Effect Magnitude.................9

Other Indices of Effect Magnitude..............................10

The Method of Meta-Analysis....................................11

The Statistical Analysis of Effect Size Data...................13

Assumptions and the Statistical Model..........................13

Estimating Effect Size.........................................14

Figure 1.......................................................16

An Unbiased Estimator of Effect Size...........................17

The Asymptotic Distribution of the Unbiased Estimator..........18

Testing Homogeneity of Effect Size.............................19

Assessing Variability of Effect Sizes..........................21

Table 1........................................................23

Figure 2.......................................................24

Estimation of Effect Size from a Series of Homogeneous Studies........25

Efficiency of the Weighted Estimator...........................28

An Analogue to the Analysis of Variance for Effect Sizes.......29

Testing Homogeneity across Classes.............................32

Testing Homogeneity of Effect Sizes within Classes.............33

An Analogy to the Analysis of Variance.........................34

Computational Formulas for $H_T$, $H_B$, and $H_W$....................35

Fitting Effect Size Models to a Series of Studies..............36

Comparisons between Classes....................................38

An Analogue to Multiple Regression for Effect Sizes............40

Testing Model Specification....................................46

Table of Contents continued

Computing Estimates and Test Statistics..............................45

The Effects of Measurement Error on Effect Size..................50

The Effects of Departures from Linear Equatability..............52

Statistical Analysis When Correlations or Proportions

are the Index of Effect Magnitude..........................55

Statistical Analysis for Correlations as Effect Magnitudes.......56

Statistical Analysis for Differences in Proportions.............61

Conclusions..................................................65

References...................................................71

## INTRODUCTION

The educational research enterprise has grown tremendously in the last thirty years. The literature in many areas of education and psychology has produced hundreds of studies on the same topic. Yet few would argue that the knowledge base of the social sciences has grown as rapidly as the volume of research studies. Some critics and many reviewers contend that our state of knowledge has remained unchanged despite the best efforts of the social science research community. Until recently, research reviews that yield equivocal conclusions have been the rule rather than the exception. Glass (1976) noted that "the typical reviewer concludes that the research is in horrible shape; sometimes one gets results, sometimes one does).

The recurrence of equivocal conclusions from research reviews led some investigators to speculate that the <u>process</u> of research review might be at fault. Light and Smith (1971) were among the first investigators to examine the problem of integrating the results of quantitative studies in the social sciences. They demonstrated the importance of systematic analysis of variations in design and execution of studies as well as the variation in study outcomes.

Light and Smith also generalized an approach from cluster sampling to generate an extensive algorithm and analysis strategy for a series of similar experiments: Unfortunately, their approach requires access to the original data which limits its practical usefulness in research integration.

Light and Smith asserted that, at that time, a technique called vote-counting was the most commonly used method of integrating research studies. In their formulation, a number of studies compare the soores of tes of two groups; one group of subjects receives an experimental treatment and the other group receives no treatment. In the vote-counting method the available studies are sorted into three categories: those that yield positive significant results, those that yield negative significant results, and those that yield nonsignificant results.

> If a plurality of studies falls into any of these three categories, with fewer falling into the other two, the modal category is declared the winner. This modal categorization is then assumed to give the best estimate of the true relationship between the independent and dependent variables. (Light & Smith, 1971, p. 433).

Despite the obvious simplicity of vote-counting methods, these techniques have very serious problems. The deficiency of vote-counting methods stems from their reliance on tests of statistical significance in individual research studies. Hedges and Olkin (1980) proved that when studies typically use smly use small samples or when the phenomenon under study produces small effects, vote-counting methods systematically fail to detect effects. The reason for this behavior is related to the low statistical power of significance tests when effects or sample sizes are small. Small effects are the rule rather than the exception in social science research.

For example, Gage (1978) has noted that the magnitude of the relationship between any teaching variable and achievement is likely to be small, although the cumulative effect of many such variables need not be negligible. Similar arguments have been made about the magnitude of relationships in social psychology.

The consequence of small effects and sample sizes on the power of statistical analyses in educational and psychological research is illustrated in surveys of statistical power of published research. Brewer (1972) calculated the power of studies published in three educational research journals. His analysis showed that the analysis showed that the power of published studies to detect small effects (a mean difference of 0.2 in standard deviation units) was uniformly low. Only two per cent of the 55 studies surveyed from the American Educational Research Journal had a power greater than 0.3 to detect an effect that small. Thus the probability of Type II errors (i.e., failure to reject the null hypothesis when it is false) seems unacceptably high in these studies. Similar results have been found in surveys of studies in abnormal psychology (Cohen, 1962), communication research (Katzer & Sodt, 1973), and applied psychology (Chase & Chase, 1976). If these surveys of social science research are representative, failure to reject the null hypothesis in individual research studies cannot provide much assurance that small effects are not present.

A new approach to the problem of research integration was proposed by Glass (1976). He argued that estimation of the magnitude of the experimentperimental effect is perhaps more important than statistical significance. Glass suggested that the "effect size" in a, two-group experiment be defined as the difference between the experimental and control group means divided by the control group standard deviation. Glass coined the term "meta-analysis" to describe the analysis of these "effect sizes" from a series of studies.

Meta-analysis has become an important supplement to traditional methods of research reviewing, largely as a result of the work of Glass and his colleagues. They demonstrated that the technique could be used to provide sensible answers to fundamental questions in the behavioral sciences. The first application of meta-anaylysis was the integration of studies on the effects of psychotherapy (Smith & Glass, 1977). This first meta-analysis intrigued many and stirred controversy for others. A series of other analyses, including the meta-analyses of the effects of class-size (Glass & Smith, 1979; Smith & Glass, lith & Glass, 1980). have continued to provide strong evidence on long standing controversies. The interest generated by these and other examples, along with a lucid treatment of the methods of meta-analysis (Glass, 1978) have encouraged other investigators to use the technique.

Many problems must be addressed by the reviewer who carries out a meta-analysis. These problems include identifying and obtaining appropriate studies, extracting estimates of effect size from the studies, coding or classifying studies, analyzing the data, and reporting the results of the data analysis. Some of these problems are similar to the problems faced by the primary researcher (see Jackson, 1980, or Cooper, 1982). In other cases the problems are not the same as those faced in primary research. The best source on methods for conducting meta-analyses is the book by Glass, McGaw, and Smith (1981). This book contains opularizing the method of meta-analysis.

Why then another paper on a reasonably complete coverage of methods to deal with all of the problems mentioned above as well as numerous others. This book draws upon the considerable resources of three authors who have been instrumental in developing and popularizing the method of meta-analysis.

Why then another paper on methodology for meta-analysis? Since the publication of the Glass, McGaw, and Smith book, there has been a great deal of interest in the development of systematic statistical theory for meta-analysis. Many of the techniques proposed and used by Glass and his associates were sensible, but suboptimal. Recent work in the statistical theory for meta-analysis has provided simple methods that can be rigorously justified. The purpose of this monograph is to supplement the existing literature on meta-analysis by providing a unified treatment of rigorous statistical methods for meta-analysis.

## Indices of Effect Size

Statistical methods have been used to combine information from different research studies for many years. Some of the earliest examples of this work are found in the work on combining the results of agricultural experiments. Cochran (1937) considered the problem of combining estimates of treatment effects from a series of similar experiments. He considered several methods of weighting estimates from each experiment. Yates and Cochran (1938) developed more refined weighting methods (e.g., partial weighting) for combining estimates from several agricultural experiments. A more recent review of statistical work in this tradition is given in Cochran (1954). Tests of the statistical significance of combined results were also introduced in connection with problems of combining results of several studies.in agriculture and biology (e.g., Tippett, 1931; Pearson, 1933).

The early work on combining the results of studies in agriculture involved combining the results of studies that share a common, well-defined dependent variable. For example, the object of a research synthesis in agriculture might be to combine estimates of the barley crop.yield derived from several studies. Each study would measure the dependent variable in the same way: the number of pounds of barley yielded per acre planted. Therefore the means or treatment effect estimates derived as mean differences are directly comparable and can be directly combined by averaging. When a series of studies in the social sciences use the same measure of the dependent variable, methods developed for combining the results of agricultural experiments can be used to combine estimates of the treatment effect.

For many common educational variables, a variety of different psychological tests provide reasonably adequate measures of the underlying construct. Some authors (e.g., Campbell, 1969) have argued that the different measures (operationalizations) of a construct are highly desirable. Studies of different operationalizations of a construct allow researchers to "triangulate" on the construct. Given a series of tests that are supposed to measure the same construct, one might ask what is meant by "measure the same construct." One definition is that two tests measure the same construct if true scores on the tests are perfectly correlated. This implies that the tests are measures of the same construct if they are linearly equatable except for errors of measurement. This notion is the basis for deciding that two tests are measuring the same thing if they yield intercorrelations that are about as high as their reliabilities will allow.

## The Effect Size as an Index of Effect Magnitude

Glass (1976) suggested the standardized mean difference or effect size as the scale invariant measure of treatment effect. We define the effect size $\delta$ for an experiment as

$$\delta = \frac{\mu^E - \mu^C}{\sigma} .$$

where $\mu^E$ and $\mu^C$ are the experimental and control group population means and $\sigma$ is the within-group population standard deviation. If the same experiment had been performed using a different (linearly equatable) measure of the outcome variable, the effect size would not change. The effect size is invariant under linear transformations of outcome variables. Therefore the effect size provides an index of effect magnitude that is independent of the particular test used to measure a construct.

We emphasize that the effect size is only invariant under substitution of (linearly equatable) measures of the same construct. Effect sizes are not invariant under nonlinear rescaling. Similarly, there is no reason to believe that effect sizes derived from measures of one construct are equivalent to effect sizes derived from measures of another construct. Different constructs will, in general, yield different effect sizes. Thus the notion of effect size of a treatment should really be considered as effect size of a treatment on a construct.

## Other Indices of Effect Magnitude

Glass (1978) also observed that the product moment correlation is a scale invariant measure of the relationship between two continuous variables. That is, the correlation does not change as a result of linear rescalings of variables. He therefore suggested that correlation coefficients could be used as indices of effect magnitude for studies examining the relationship between two continuous variables. In some cases, the natural index of effect magnitude is the difference between proportions of subjects that reach a criterion in experimental and control groups. The proportions are themselves scale invariant and can therefore be used directly to compute a scale invariant index of effect magnitude such as the difference between proportions.

## The Method of Meta-Analysis

The object of meta-analysis or other methods of quantitative research synthesis is to use data from a series of studies to obtain information about the effect size for a treatment on various constructs. This usually involves obtaining an estimate of effect size from each study and pooling (averaging) these estimates to obtain an estimate of the average effect size across studies (Glass, 1976). In addition, the investigator may want to determine whether any characteristics of the studies are systematically related to effect size.

Some writers in the area of research synthesis have cited substantive reasons for the position that different studies of the effects of the same treatment might yield quite different results. Light and Smith (1971) argued that many contradictions in research evidence may be resolved by grouping studies with similar characteristics. They asserted that studies with the same characteristics are more likely to yield similar results, and hence many apparent contradictions among research results arise from differences in the characteristics of studies. Pillemer and Light (1980) have argued that examining the relationship of variations in study outcomes and study characteristics is an essential step in assessing the range of generalizability of a research finding. For example, if a treatment produces essentially the same effect in a wide variety of settings with a variety of people, we are more confident in the generalizability of the finding of a treatment effect related to effect size. The statistical analyses have sometimes involved regressing the effect size estimates obtained from a series of studies on variables that represent various characteristics of studies (Glass, 1976, 1978). Such methods have been used, for example, in the meta-analysis of studies of the effectiveness of psychotherapy (Smith and Glass, 1977), the effects of class size on achievement (Glass and Smith, 1979), and the effects of television on achievement (Pascarella, Walberg, Junker, & Haertel, 1981).

## THE STATISTICAL ANALYSIS OF EFFECT SIZE DATA

### Assumptions and the Statistical Model

Many treatments of effect size have not adequately emphasized the assumptions underlying effect size estimation and testing. Glass (1976) proposed the quantitative synthesis of the results of a collection of experimental/control group studies by estimating a population effect size for each study and then combining the estimates across studies. The statistical analyses in such studies typically involve the use of a *t*- or *F*-test to test for differences between the groups. If the assumptions for the validity of the *t*-test are met, it is possible to derive the properties of estimators of effect size exactly. We start by stating these assumptions explicitly.

Suppose that the data arise from a series of k independent studies, where each study compares an experimental group (E) with an independent control group (C). Let $Y_{ij}^E$ and $Y_{ij}^C$ be the *j*th scores on the *i*th experiment from the experimental and control groups, respectively. Assume that for fixed *i*, $Y_{ij}^E$ and $Y_{ij}^C$ are normally distributed with means $\mu_i^E$ and $\mu_i^C$ and common variance $\sigma_i^2$, i.e.,

$$Y_{ij}^E \sim N(\mu_i^E, \sigma_i^2),\ j = 1,\ldots,n_i^E,\ i = 1,\ldots,k,$$

and

$$Y_{ij}^C \sim N(\mu_i^C, \sigma_i^2),\ j = 1,\ldots,n_i^C,\ i = 1,\ldots,k.$$

In this notation, the effect size for the *i*th study ($\delta_i$) is defined as

$$\delta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i} \qquad (1)$$

where we use the Greek letter $\delta$ to denote that this effect size is a population parameter.

Note that the assumptions of the t-test may not always be met in practice. They may never be exactly met. These assumptions are often reasonably well satisfied in practice, and the theory that follows, as well as that of the primary statistical analyses, will be a reasonable approximation to reality. Since the theory that follows relies on the properties of the t-distribution, many of the results should be robust. In some situations, however, violations of the model assumptions will be severe. For example, the observations in each study might have a highly skewed distribution. In cases such as these, alternative statistical methods are necessary. Unfortunatly, there has been little work on statistical procedures for meta-analysis with nonstandard models. One exception is that Kraemer and Andrews (1982) have provided a "nonparametric" estimator of effect size. Additional work is needed to provide a more complete theory for meta-analysis when standard assumptions are not tenable. Another important issu is the quality of the data reported in studies to be combined. The quality of the research synthesis is unlikely to be higher than that of the studies that go into it. This suggests that reviewers must carefully examine the studies before an attempt is made to combine the results of those studies.

## Estimating Effect Size

The definition of effect size given in (1) above defines a population parameter $\delta_i$ in terms of other population parameters

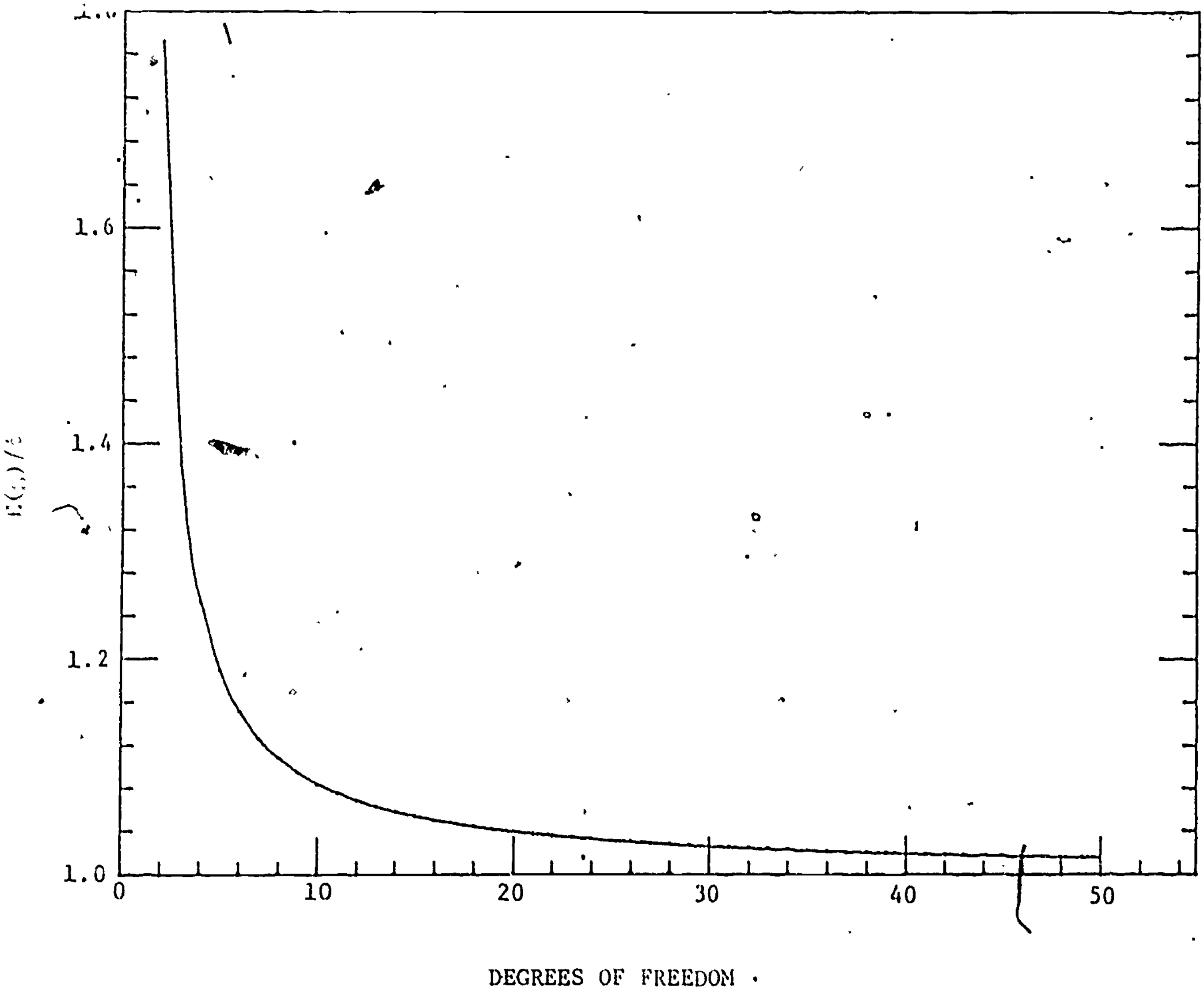
$\mu_i^E$, $\mu_i^C$, and $\sigma_i$. We will seldom, if ever, know the exact values of $\mu_i^E$, $\mu_i^C$, and $\sigma_i$, thus we will have to estimate $\delta_i$. Glass (1976) proposed a statistic $g_i'$ to estimate $\delta_i$ by essentially replacing $\mu_i^E$, $\mu_i^C$, and $\sigma_i$ in the definition of $\delta_i$ by their sample analogues. Specifically Glass proposed the estimator $g_i'$ of $\delta_i$, where $g_i'$ is defined by

$$g_i' = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{S_i^C}, \; i = 1,\ldots,k, \qquad (2)$$

where $Y_i^E$ and $Y_i^C$ are the experimental and control group sample means for the ith study .d $S_i^C$ is the control group sample standard deviation. Hedges (1981) has shown that under the assumptions of the previous section, the estimator (2) is biased. Figure 1 is a graphic representation of the relationship between the ratio of the expected value of g to the true parameter value $\delta$ as a function of the degrees of freedom in the estimate of $\sigma_i$. We see that the bias of $g_i'$ tends toward zero in studies with large sample sizes but can be substantial in studies with small sample sizes.

If the assumption of equal population variances in experimental and control groups holds, a less biased estimator results when $S_i^C$ is replaced with the usual pooled within-groups

Figure 1



The ràtio E(g)/δ of the expectation of the estimator $\tilde{g} = (\overline{Y}^E - \overline{Y}^C)/S$ to the true effect size δ as a function of m, the degrees of freedom of S used to estimate σ.

standard deviation. We denote this estimator by $\tilde{g}_i$, that is,

$$\tilde{g}_i = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{S_i}, \quad i = 1,\ldots,k, \qquad (3)$$

where $S_i^2$ is the pooled estimate of the variance

$$S_i^2 = \frac{(n_i^E - 1)(S_i^E)^2 + (n_i^C - 1)(S_i^C)^2}{n_i^E + n_i^C - 2} .$$

We emphasize that $\tilde{g}_i$ is a sample statistic and therefore has a sampling distribution of its own. Our assumptions imply that $g_i$ is distributed as $(1/\sqrt{n_i})$times a noncentral t random variable with $n_i^E + n_i^C - 2$ degrees of freedom and noncentrality parameter $\sqrt{\tilde{n}_i \delta_i}$, where $\tilde{n}_i = n_i^E n_i^C / (n_i^E + n_i^C)$. This distribution leads immediately to exact expressions for the bias and variance of $g_i$, which are given in Hedges (1981). One should also note that $g_i$ is an inference sufficient statistic for $\delta_i$.

## An Unbiased Estimator of Effect Size

A simple unbiased estimator of $\delta$ was obtained by Hedges (1981) based on the assumptions of the previous section: The unbiased estimator $g_i$ is given by

$$g_i = c(m)\tilde{g}_i, \qquad (4)$$

where $m = n_i^E + n_i^C - 2$, $c(m)$ is given exactly by

$$c(m) = \frac{\Gamma(m/2)}{\sqrt{m/2}\;\Gamma[(m-1)/2]}\,, \qquad (5)$$

$\Gamma(x)$ is the gamma function and $c(m)$ is given approximately by

$$c(m) \doteq 1 - \frac{3}{4m - 1}\,.$$

It is clear that as m becomes large, $g_i$ tends to $\tilde{g}_i$ , so that $g_i$ is almost unbiased in large samples. Since $c(m) < 1$, the variance of the unbiased estimator $g_i$ is always smaller than the variance of $\hat{g}_i$. Hence $g_i$ has uniformly smaller mean squared error that $\breve{g}_i$. The exact variance of $g_i$ is

$$\frac{[c(n_i^E + n_i^C - 2)]^2[n_i^E + n_i^C - 2][1 + \tilde{n}_i\delta^2]}{(n_i^E + n_i^C - 4)\tilde{n}_i} - \delta^2\,, \quad (6)$$

where $\check{n}_i = n_i^E n_i^C / (n_i^E + n_i^C)$, and $c(m)$ is given by (5).

## The Asymptotic Distribution of the Unbiased Estimator

In small samples, the estimator $g_i$ of effect size has a sampling distribution that is a constant times the noncentral t-distribution. When the sample sizes in the experimental and control groups are large, however, the asymptotic distribution of $g_i$ provides a satisfactory approximation to the exact distribution of $g_i$. The large sample approximation is given by

$$g_i \sim N(\delta_i, \sigma_i^2(\delta_i)), \qquad (7)$$

where

$$\sigma_i^2(\delta_i) = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{\delta_i^2}{2(n_i^E + n_i^C)} \qquad (8)$$

and we use the expression $\sigma_i^2\,(\delta_i)$ to indicate that the variance of $g_i$ depends on the true effect size $\delta_i$. This large sample approximation is used by substituting an estimator of the effect size for $\delta_i$ in (8). In the case of a single effect size, we substitute $g_i$ for $\delta_i$ in (8) to obtain an expression for the variance of $g_i$. A useful guideline on what constitutes a large sample is $n^E$, $n^C \geq 10$. If the sample size of either group is smaller than about 10, it may be desirable to omit the study from data analyses since the estimate of effect size is so imprecise that it is almost useless.

## Testing Homogeneity of Effect Size

Before pooling estimates of effect size from a series of k studies, it is important to ask whether the studies can reasonably be described as sharing a common effect size. A statistical test for the homogeneity of effect size is formally a test of the hypothesis

$$H_o: \ \delta_i = \delta, \ i = 1, \ldots, k,$$

versus the alternative that at least one $\delta_i$ differs from the rest.

A large sample (approximate) test for the equality of k effect sizes given by Hedges (1982a) uses the test statistic

$$H_T = \sum_{i=1}^{k} \frac{(g_i - g.)^2}{\sigma_i^2(g_i)} , \qquad (9)$$

where g. is the weighted estimator of effect size given below in (13).

The test statistic $H_T$ is the sum of squares of the $g_i$ about the weighted mean g., where the $\underline{i}^{th}$ square is weighted by the reciprocal of the estimated variance of $g_i$. The defining formula (9) is helpful in illustrating the intuitive nature of the statistic $H_T$, but a computational formula is more useful for actual calculation of $H_T$. The computational formula is

$$H_T = \sum_{i=1}^{k} \frac{(g_i)^2}{\sigma_i^2(g_i)} - \frac{\left(\sum_{i=1}^{k} \frac{g_i}{\sigma_i^2(g_i)}\right)^2}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2(g_i)}} , \qquad (10)$$

where $\sigma_i^2(\delta_i)$ is given by (8). A similar test is given by Rosenthal and Rubin (1982).

When each study has a large sample size (a reasonable guideline is $n^E$, $n^C \geq 10$), the asymptotic distribution of $H_T$ can be used as the basis for an approximate test of the homogeneity of the $\delta_i$. If all the k studies have the same population effect size (i.e., if $H_0$ is true) then the test statistic $H_T$ has an asymptotic chi-square distribution given by

$$H_T \sim \chi^2_{k-1} .$$

Therefore if the obtained value of $H_T$ exceeds the 100(1-α) per cent critical value of the chi-square distribution with (k-1) degrees of freedom, we reject the hypothesis that the $\delta_i$ are equal. If this null hypothesis is rejected, a conservative individual may decide not to pool all of the estimates of δ since they are not estimating the same parameter. When the sample sizes are very large, however, it is probably worthwhile to consider the actual variation in the values of $g_i$, since rather small differences may lead to large values of the test statistic. If the $g_i$ values do not differ much in an absolute sense, the investigator may elect to pool the estimates even though there is reason to believe that the underlying parameters are not identical. A less conservative investigator might pool estimates regardless of the outcome of tests of homogeneity.

## Assessing Variability of Effect Sizes

It is often helpful to plot the effect sizes from a series of studies to assess the variability of the $g_i$ values. The large sample approximation (7) may be used to obtain a confidence interval for each effect size. An approximate $100(1-\alpha)$ per cent confidence interval for $\delta_i$ is given by

$$g_i - z_{\alpha/2}\sigma_i(g_i) \le g_i + z_{\alpha/2}\sigma_i(g_i) ,$$

where $z_\alpha$ is the $100\alpha$ per cent critical value of the standard normal distribution and $\sigma_i^2(g_i)$ is the large sample variance of $g_i$ given by (8). Plotting each $g_i$ value along with a confidence interval for each $g_i$, gives an idea of the region in which the corresponding $\delta_i$ is likely to be. Therefore substantial overlap of these confidence intervals suggests that there is agreement among the $g_i$ on a common effect size. Conversely, if some of the $g_i$ values are far from the rest, and their associated confidence intervals do not overlap much, then it may be useful to consider these deviant values as outliers. If there are only a few outlying values then it may be helpful to treat these studies separately, and estimate a common effect size from the other studies.

The effect sizes from 10 studies of the effect of open education on attitude toward school are presented in Table 1 along with sample sizes and the estimated sampling standard deviation $\sigma_i(g_i)$ for each study. The 95 per cent confidence intervals for these effect sizes are plotted in Figure 2. We see that one effect size, that of study 10, is quite a bit larger than the rest. Similarly, the confidence interval for $\delta_{10}$ fails to overlap with those of other studies. Calculations for the test of homogeneity are also given in Table 1, and we see that the value of the homogeneity statistic $H_T = 19.40$ which is

Table 1

Effect Sizes from 10 Studies of the Effects of Open Education on Student Attitude toward School

| Study | $n^E$ | $n^C$ | $g$ | $\sigma^2(g)$ | $1/\sigma^2(g)$ | $g/\sigma^2(g)$ | $g^2/\sigma^2(g)$ |
|---|---|---|---|---|---|---|---|
| 1 | 131 | 138 | .158 | .0149 | 66.996 | 10.585 | 1.672 |
| 2 | 40 | 40 | .261 | .0504 | 19.831 | 5.176 | 1.351 |
| 3 | 40 | 40 | .649 | .0526 | 19.000 | 12.331 | 8.003 |
| 4 | 79 | 49 | .503 | .0341 | 29.365 | 14.770 | 7.429 |
| 5 | 84 | 45 | .458 | .0349 | 28.620 | 13.108 | 6.004 |
| 6 | 78 | 55 | .577 | .0322 | 31.004 | 17.889 | 10.322 |
| 7 | 38 | 110 | .588 | .0366 | 27.341 | 16.077 | 9.453 |
| 8 | 38 | 93 | .392 | .0376 | 26.557 | 10.410 | 4.081 |
| 9 | 20 | 23 | -.055 | .0935 | 10.694 | -.588 | 0.032 |
| 10 | 40 | 40 | -.332 | .0507 | 19.728 | -6.550 | 2.175 |
| | | TOTALS | | | 279.135 | 93.209 | 50.522 |

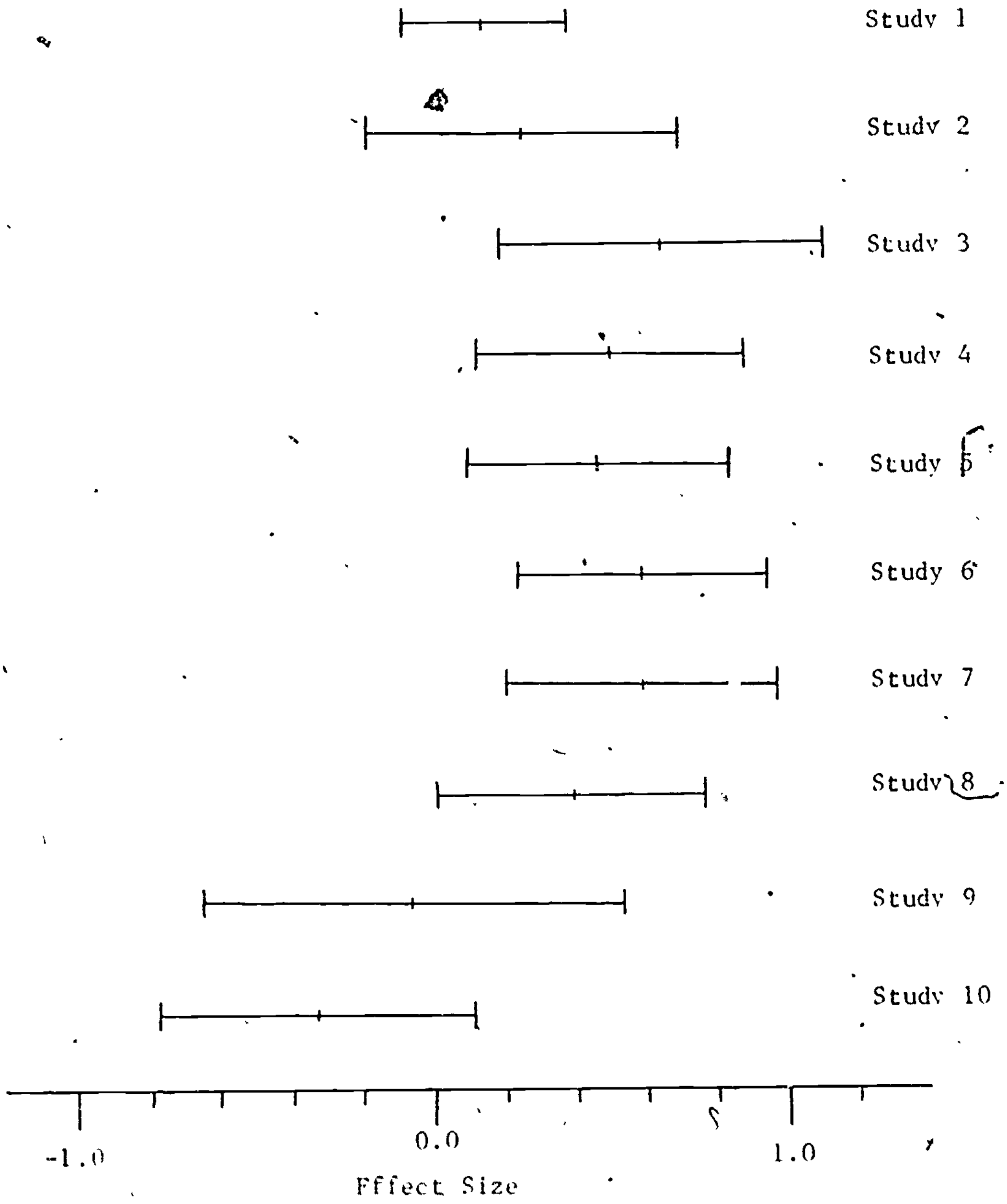


Figure 2. Ninety-five per cent confidence intervals for effect sizes for the ten studies described in Table 1.

significant beyond the $\alpha = .025$ level. Deleting the effect size for study 10, we see that the effect sizes are reasonably homogeneous: $H_T = 9.983$, $.10 \leq p \leq .05$.

Estimation of Effect Size from a Series of Homogeneous Studies

If a series of k independent studies share a common effect size $\delta$, it is natural to estimate $\delta$ by pooling estimates from each of the studies. If the sample sizes of the studies differ, then the estimates from some (the larger) studies will be more precise than the estimates from other (smaller) studies. In this case it is reasonable to give more weight to the more precise estimates when pooling. This leads to weighted estimators of the form

$$\sum_{i=1}^{k} w_i g_i, \qquad (11)$$

where $w_i > 0$, $\bar{i} = 1,\ldots,k$, and $\sum_{i=1}^{k} w_i = 1$. It is easy to show that the weights that minimize the variance of (11) are given by

$$w_i = \frac{1/v_i}{\sum_{j=1}^{k} 1/v_j}, \; i = 1,\ldots,k, \qquad (12)$$

where $v_i$ is the variance of $g_i$ given in (6). The practical problem in calculating the most precise weighted estimate is that the $\underline{i}^{th}$ weight depends on the variance of $g_i$ which in turn depends on $\delta$.

One approach to the problem of weighting results from different studies, is to use weights that are based on some approximation to the $v_i$ that does not depend on $\delta$. This procedure results in a pooled estimator that is unbiased, but it will usually be less precise than if the optimal weights are used. For example, weights could be derived by assuming that

$$v_i = [c(n_i^E + n_i^C - 2)]^2(n_i^E + n_i^C - 2)/\tilde{n}_i(n_i^E + n_i^C - 4).$$

The weights thus derived are only optimal if $\delta = 0$. If $\delta$ is near zero these weights will be close to optimal since $v_i$ depends on $\delta^2$, which will be small. If a nonzero a priori estimate of $\delta$ is available, then weights could be estimated by inserting that value of $\delta$ in expression (6) for the variance of $g_i$ and using the formula (12) for $w_i$. In general the result will be an unbiased pooled estimator of $\delta$ that is slightly less precise than the most precise weighted estimator.

Another approach to obtaining a weighted estimator of $\delta$ is to estimate $\delta$ and use the sample estimate of $\delta$ to estimate the weights for each study. Define the weighted estimator g. by

$$g_{\cdot} = \frac{\sum_{i=1}^{k} \frac{g_i}{\sigma_i^2(g_i)}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2(g_i)}}, \qquad (13)$$

where $\sigma_i^2(g_i)$ is given by (8). The estimator g. is therefore obtained by calculating the weights using $g_i$ for $\delta_i$ in (8). Although the $g_i$ are unbiased, g. is not. The bias of g. is small in large samples and tends to zero as the sample sizes tend to infinity.

This estimator could be modified by replacing $g_i$ by g. in the expression for $\sigma_i^2(g_i)$, and iterating. That is, calculate the estimator $g_{\cdot}^{(1)}$ defined by

$$g_{\cdot}^{(1)} = \frac{\sum_{i=1}^{k} \frac{g_i}{\sigma_i^2(g_\cdot)}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2(g_\cdot)}}, \qquad (14)$$

where $\sigma_i^2(\delta_i)$ is given by (8). The iterated estimator $g_{\cdot}^{(1)}$ will tend to be less biased than g. . If the effect size is homogeneous across experiments, the iteration process usually will not change the estimate very much.

The asymptotic distribution of g. is easily obtained and can be used to obtain large sample confidence intervals for $\delta$ based on g. . The formal definition of 'large sample' in this case is that the sample sizes $n_i^E$ and $n_i^C$, $i = 1,\ldots,k$ are tending to infinity at the same rate. A practical guideline for 'large sample' is $n^E, n^C \geq 10$. The large sample approximation is

$$g_\cdot \sim N(\delta, \sigma_\cdot^2(\delta)), \qquad (15)$$

where

$$\sigma_\cdot^2(\delta) = \frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma^2(\delta)}}, \qquad (16)$$

and $\sigma^2(\delta)$ is given by (8). We use this large sample approximation by substituting the (consistent) estimator g. for $\delta$ in (15). A 100(1 $-\alpha$) per cent asymptotic confidence interval for $\delta$ is therefore

$$g. - z_{\alpha/2}\sigma.(g.) \leq \delta \leq g. + z_{\alpha/2}\sigma.(g.),$$

where $z_{\alpha/2}$ is obtained from a table of the standard normal distribution. Similarly, an asymptotic test of the hypothesis that $\delta = 0$ uses the test statistic

$$z(g.) = \frac{g.}{\sigma.(g.)} . \qquad (17)$$

If the obtained value of z(g.) is larger in absolute value than the 100(1 $- \alpha/2$) per cent critical value of the standard normal distribution, we reject the hypothesis that $\delta = 0$ at the 100$\alpha$ per cent significance level.

The formal asymptotic distribution of the iterated estimator $g.^{(1)}$ is the same as that of g. . We use the large sample approximation to the distribution of $g.^{(1)}$ by substituting $g.^{(1)}$ for $\delta$ in (16). Therefore confidence intervals and significance tests for $\delta$ based on $g.^{(1)}$ are calculated in the same way as for g. . The only difference when using $g.^{(1)}$ is that g. is replaced by $g.^{(1)}$ wherever the former occurs.

## Efficiency of the Weighted Estimator

The weighted estimators discussed in previous sections were derived by finding the expression for weights that minimize the variance of the resulting weighted estimator. One might ask whether the best (most precise) weighted estimator is the most precise in some larger class of estimators of effect size, including those that are not weighted linear combinations of the $g_i$. Hedges (1982a) showed that g. is asymptotically efficient in the sense that the asymptotic variance of g. is the theoretical minimum (Cramér-Rao bound). Thus no other consistent estimator has smaller asymptotic variance. This result implies that g. has the same asymptotic distribution as the maximum likelihood estimator of $\delta$ based on k experiments.

## An Analogue to the Analysis of Variance for Effect Sizes

The representation of the results of a collection of studies by a single estimate of effect magnitude can be misleading if the underlying (population) effect sizes are not identical in all of the studies. For example, suppose a treatment produces large positive (population) effects in one-half of a collection of studies, and large negative (population) effects in the other half of a collection of studies. Then representation of the overall effect of the treatment as zero is misleading, because all of the studies actually have underlying effects that are different from zero. The test for homogeneity of effect size given in (9) provides a method for detecting heterogeneity of effect sizes. It will often be the case that a collection of studies cannot be reasonably said to share the same effect size. For example, Giaconia and Hedges (1982) report the results of tests of homogeneity for studies measuring the effect of open education on 19 different dependent variables. For each dependent variable, the hypothesis of homogeneity of effect size was easily rejected.

Some investigators in quantitative research synthesis (e.g., Kulik, Kulik, & Cohen, 1979) have recognized the potential for heterogeneous effect sizes and have grouped studies which share common characteristics into classes. The usual approach is then to treat the effect size estimates as data and calculate an analysis of variance to determine if these classes have different mean effect sizes. There are two problems with this procedure. First, the assumptions of the analysis of variance may not be met since the effect size estimates may not have the same distribution within cells. The variance of an individual observation (effect size estimate) is proportional to 1/n, where n is the number of subjects in the study. When studies have different sample sizes, the individual "error" variances may differ by a factor of 10 or 20. Secondly, even if the between-classes test were accurate, the use of ANOVA does not provide any indication whether or not studies _within the classes_ share a common effect size. Thus, even if ANOVA correctly detects that two classes of studies have a different average effect size, there is no guarantee that the average effect size within each class is a reflection of a common underlying effect size for that class.

Hedges (1982b) presented an alternative technique for fitting models to effect sizes from a series of studies. We assume the investigator has an _a priori_ grouping of studies, that is, a scheme for classifying studies that are likely to produce similar results. This will often take the form of a set of categories into which studies may be placed. Studies may be cross classified by two or more sets of categories. The technique presented in this section is straightforward. Conceptually the investigator begins by asking whether all studies (regardless of category) share a common effect size. A statistical test (fit statistic) is provided by the test of homogeneity given in (9). If the hypothesis of fit to a single effect size is rejected, the experimenter then breaks the series of studies into classes, and asks whether the model of a different effect size of each class fits the data. It is interesting to note that the fit statistic calculated at the first stage is partitioned into stochastically independent parts corresponding to between-class and within-class fit, respectively. The between-class fit is an index of the extent to which effect sizes in the classes are different. If the within-class fit (fit to a single effect size within each class) is not rejected, the investigator may stop. If the within-class fit is rejected, the investigator may want to further subdivide the classes. The process of subdividing and testing for between- and within-class fit continues until an acceptable level of within-class homogeneity is achieved. The procedure provides valid asymptotic tests for the effects of classifications as well as an indication that the final classes are internally homogeneous with respect to effect size.

Testing Homogeneity across Classes

Suppose that the entire collection of studies is divided into p *a priori* classes. The test for homogeneity across classes is essentially a test that the average effect size in each class is the same as the average effect size in every other class. Hedges (1982b) gave the test statistic $H_B$ to test for homogeneity of effect size across classes. The statistic $H_B$ is given by

$$H_B = \Sigma_{j=1}^{p} \Sigma_{i\varepsilon Ij} \frac{(g_{j.} - g_{..})^2}{\sigma_i^2(g_i)} , \qquad (18)$$

where $\Sigma_{i\varepsilon Ij}$ is the sum over all studies with subscript *i* in the *j*th class, $g_{j.}$ is the weighted average effect size for the *j*th class given by

$$g_{j.} = \frac{\Sigma_{i\varepsilon Ij} \frac{g_i}{\sigma_i^2(g_i)}}{\Sigma_{i\varepsilon Ij} \frac{1}{\sigma_i^2(g_i)}} , \qquad (19)$$

$g_{..}$ is the weighted average effect size based on all of the studies given by (13) or alternatively

$$g_{..} = \frac{\Sigma_{j=1}^{p} \Sigma_{i\varepsilon Ij} \frac{g_i}{\sigma_i^2(g_i)}}{\Sigma_{j=1}^{p} \Sigma_{i\varepsilon Ij} \frac{1}{\sigma_i^2(g_i)}} , \qquad (20)$$

and $\sigma_i^2(g_i)$ is given in (8).

If the effect sizes are identical in each class, then the test statistic $H_B$ given in (18) has an asymptotic distribution given by

$$H_B \sim \chi^2_{p-1} \qquad (21)$$

Therefore the test of homogeneity of effect size across classes at a significance level $\alpha$ consists of comparing the obtained value of $H_B$ with the 100(1 $-\alpha$) per cent critical value of the chi-square distribution with (p − 1) degrees of freedom. If $H_B$ is greater than the critical value, we reject the hypothesis of homogeneity of effect size across classes.

Testing Homogeneity of Effect Sizes within Classes

The test of homogeneity of effect size within classes is a test whether all of the effect sizes within the same class share a common effect size. Hedges (1982b) gave the test statistic $H_W$ for testing the homogeneity of effect size within classes. This test statistic is the sum of the test statistics $H_{Wj}$ for the homogeneity of effect size within the $j^{th}$ class. Thus the statistic $H_W$ is given by

$$H_W = \Sigma_{j=1}^{p} \sum_{i \varepsilon Ij} \frac{(g_{j.} - g_i)^2}{\sigma_i^2(g_i)}, \qquad (22)$$

where $\sum_{i \varepsilon Ij}$ and $g$ are defined as in (19) and $\sigma_i^2(g_i)$ is given in (8). Alternatively we could calculate $H_W$ as

$$H_W = \Sigma_{j=1}^{p} H_{wj},$$

where

$$H_{Wj} = \Sigma_{i\varepsilon Ij} \frac{(g_{j.} - g_i)^2}{\sigma_i^2(g_i)}, \; j = 1,\ldots,p.$$

If the effect sizes within each class are homogeneous, then $H_W$ has an asymptotic distribution given by

$$H_W \sim \chi^2_{k-p}. \qquad (23)$$

Therefore the test for homogeneity of effect sizes within classes consists of comparing the obtained value of $H_W$ with the 100(1 - α) per cent critical value of the chi-square distribution on (k - p) degrees of freedom. If the obtained value of $H_W$ exceeds the critical value we reject the hypothesis that the effect sizes are homogeneous within classes. In data analyses, it may be helpful to calculate the within-class fit statistics $H_{Wj}$ for each of the p classes. This may facilitate the identification of classes in which the fit is particularly bad.

## An Analogy to the Analysis of Variance

There is a simple relationship among the fit statistics $H_B$, $H_W$, and $H_T$ that is analogous to the partitioning of sums of squares in the analysis of variance. It is possible to show that

$$H_T = H_B + H_W,$$

using only elementary algebra. One interpretation of this formula involves this partitioning of the fit statistic $H_T$. The "tot.l fit" to the model of a single effect size is represented by $H_T$. The "between-class fit" is represented by $H_B$ and the "within-class fit" is represented by $H_W$. Thus the total fit is partitioned into between-class and within-class components. We have stated that the statistics $H_B$, $H_W$, and $H_T$ are distributed asymptotically as central chi-squares under appropriate null hypotheses with distributions given by

$$H_T \sim \chi^2_{k-1} ,$$

$$H_B \sim \chi^2_{p-1} ,$$

$$H_W \sim \chi^2_{k-p} .$$

Furthermore, Hedges (1982b) has shown that $H_B$ and $H_W$ are asymptotically independent. Therefore the tests for between- and within-class fit are asymptotically independent.

## Computational Formulas for $H_T$, $H_B$, and $H_W$

In practice, computational formulas can simplify calculation of the fit statistics $H_T$, $H_B$, and $H_W$. These formulas are much like the computational formulas in the analysis of variance. The computational formulas permit the researcher to compute each of the fit statistics in a single pass through the data with a packaged computer program. Each of the formulas can be verified by direct algebraic manipulation. The computational formula for $H_T$ is given in (10), but is repeated here in different notation for reference.

$$H_T = \Sigma_{j=1}^{p} \Sigma_{i\epsilon Ij} \frac{g_i^2}{\sigma_i^2(g_i)} - \frac{\left(\Sigma_{j=1}^{p} \Sigma_{i\epsilon Ij} \frac{g_i}{\sigma_i^2(g_i)}\right)^2}{\Sigma_{j=1}^{p} \Sigma_{i\epsilon Ij} \frac{1}{\sigma_i^2(g_i)}},$$

$$H_{Wj} = \Sigma_{i\epsilon Ij} \frac{g_i^2}{\sigma_i^2(g_i)} - \frac{\left(\Sigma_{i\epsilon Ij} \frac{g_i}{\sigma_i^2(g_i)}\right)^2}{\Sigma_{i\epsilon Ij} \frac{1}{\sigma_i^2(g_i)}}, \quad j = 1,\ldots,p,$$

$$H_W = \Sigma_{j=1}^{p} H_{Wi},$$

$$H_B = H_T - H_W,$$

where $\Sigma_{i\epsilon Ij}$ is defined as in (18) and $\sigma_i^2(g_i)$ is given in (8).

## Fitting Effect Size Models to a Series of Studies

The statistical results of this paper can be used as part of a general strategy for fitting models to the effect sizes from a series of studies. Start witha series of studies where each study assesses the effect of a particular treatment via a two group experimental group/control group design. Suppose that the dependent variables measure the same construct and are (approximately) linearly equatable. We assume that the studies are classified according to one of the classification dimensions. The classes obtained by one partitioning may be further partitioned according to a second classification dimension, and in turn partitioned according to other dimensions.

One strategy for fitting models to effect sizes for each class is analogous to the strategy used to fit hierarchical log-linear models to contingency tables. The strategy can be described as follows.

Step 1. Ignore the classifications and fit the model of a single effect size to all the studies. The estimate of this single effect size is g.. given by (20). Calculate the fit statistic $H_T$. If the value of $H_T$ is not large or is statistically insignificant at some preset $\alpha$ level, the investigator may stop, concluding that the model of a single effect size fits the data adequately. The asymptotic distribution of g.. may be used to calculate an asymptotic confidence interval for $\delta$. If the fit statistic $H_T$ is large or statistically significant, go on to Step 2.

Step 2. A large value of the fit statistic $H_T$ indicates that effect sizes are not homogeneous across all studies, so partition the studies into classes along one dimension. One should choose the most important dimension first, that is, the dimension believed to be most related to effect size.

Calculate the between-class fit statistic $H_B$ and the within-class fit statistic $H_W$. If the value of the within-class fit statistic $H_W$ is small or is statistically insignificant, the investigator may stop, since the model of a different effect size for each class is consistent with the data. In this case, $g_{j.}$ given in (19) is the estimate of effect size for the j th class and $H_B$ represents the extent to which the effect sizes differ among classes. If $H_W$ is large or statistically significant, then go on to Step 3.

Step 3. A large value of the fit statistic $H_W$ indicates that effect sizes are not homogeneous within classes. At this point it may be useful to partition within-class fit $H_W$ into p (if there are p classes) statistics $H_{Wj}$ j = 1,...,p, where $H_{Wj}$ indicates the fit within the $j^{th}$ class. Examining the values of $H_{Wj}$ may help identify classes with especially poor fit, that is, classes in which the effect sizes are heterogeneous. This may lead the investigator to exclude some classes or studies from further analyses. Examination of within-class fit may also suggest which other classification dimensions are useful. Go on to Step 4.

Step 4. Partition the existing classes according to a second classification dimension. Repeat Step 2, that is, calculate the between- and within-class fit statistics $H_B$ and $H_W$. Proceed through Steps 2, 3, and 4 until an acceptable level of within-class fit is obtained or the classification dimensions are exhausted.

The procedure given is a practical method involving relatively simple calculations. It has the advantage that fit to the model can be assessed at each stage and it also provides a test of the relationship between the classification dimension and effect size.

## Comparisons between Classes

If a priori knowledge or a formal hypothesis test (significant value of $H_B$) lead an investigator to believe that the effect sizes are not homogeneous across classes, the investigator may wish to compare the effect sizes of different classes. More generally, the investigator may wish to test hypotheses about linear combinations of the effect sizes for the classes. Such comparisons are analogous to contracts in the analysis of variance.

The general comparison is a linear combination of the $g_{j\cdot}$ of the form

$$C_g = \Sigma_{j=1}^{p} c_j g_{j\cdot}, \qquad (24)$$

where the $c_j$, $j = 1,\ldots,p$ are known constants. In the case of a comparison between two classes, for example one of the $c_j$ might be +1, another might be -1, while the remainder might be zero. The comparison $C_g$ given in (24) may be considered an estimate of

$$C_\delta = \Sigma_{j-1}^{p} c_j \bar{\delta}_{j\cdot}, \qquad (25)$$

where $\bar{\delta}_{j\cdot}$ is the weighted average population effect size in the $j^{th}$ class given by

$$\bar{\delta}_{j\cdot} = \frac{\sum_{i\varepsilon Ij} \frac{\delta_i}{\sigma_i^2(\delta_i)}}{\sum_{i\varepsilon Ij} \frac{1}{\sigma_i^2(\delta_i)}} \qquad (26)$$

Such comparisons are easiest to interpret when effect sizes are homogeneous within classes since then $\overline{\delta}_{j.}$ is simply the (common) effect size for the studies in the $j^{th}$ class.

Hedges (1982b) used the asymptotic distribution of $g_{j.}$ to obtain the large sample approximation to the distribution of $C_g$, specifically

$$C_g \sim N(C_\delta, \sigma_c^2),$$

where $\sigma_c^2$ is estimated by

$$\hat{\sigma}_c^2 = \Sigma_{j=1}^{p} \frac{c_i^2}{\Sigma_{i \varepsilon Ij} \frac{1}{\sigma_i^2(g_i)}} . \tag{27}$$

Therefore an approximate $100(1 - \alpha)$ per cent confidence interval for $C_\delta$ is given by

$$C_g - z_{\alpha/2}\hat{\sigma}_c \leq C_\delta \leq C_g + z_{\alpha/2}\hat{\sigma}_c \quad .$$

## An Analogue to Multiple Regression for Effect Sizes

When effect sizes are heterogeneous across a series of studies, one strategy is to relate discrete characteristics of studies to effect size perhaps by using the method given in previous sections. Another procedure is the is the application of regression analysis to the estimates of effect size. Glass (1978) recommended the general strategy of coding the characteristics of studies as a vector of predictor variables and then regressing the effect size estimate on the predictors to determine the relationship between characteristics of studies and effect size. For example, Smith and Glass (1977) used linear regression to determine the relationship between several coded characteristics of studies (e.g., type of therapy, duration of treatment, internal validity of the study) and the effect size in their me.a-analysis of psychotherapy outcome studies. The same method has been used in many research syntheses, including a series of meta-analyses conducted by Walberg and his associates (e.g., Uguroglu & Walberg, 1979; Pascarella, Walberg, Junker, & Haertel, 1981). This strategy has been used in some very novel and creative ways in some research syntheses. The potential of multiple regression methods in research synthesis is perhaps best illustrated by the meta-analyses of the effects of class size (Glass & Smith, 1979; Smith & Glass, 1980).

Although the regression method advocated by Glass is appealing, there are at least two problems with the method. First, the assumptions of regression analysis are not met since the variances of the individual effect size estimates are proportional to 1/n, where n is the sample size of the study. Thus when the studies to be integrated have different sample sizes, the individual "error" variances may be dramatically different. Secondly, even if the regression coefficients are properly estimated, Glass's method gives no indication of the goodness of fit of the regression model. That is, there is no indication that the model is correctly specified.

Hedges (1982c) developed alternative methods for fitting models to effect size data when those models include continuous or discrete independent variables. These methods provide consistent, asymptotically efficient estimates of the parameters of the model and also permit large sample tests of significance. In addition the methods can provide an explicit test of the specification of the model. Thus it is possible to test whether or not a model adequately explains the observed variability in effect size estimates.

In this analysis, assume that the standardized mean difference $\delta_i$ for the $\underline{i}^{th}$ experiment depends on a vector of p fixed concomitant variables $(x_{i1}, x_{i2},\ldots,x_{ip})'$, where $p \leq k$. The vectors $(x_{i1},\ldots,x_{ip})'$, $i = 1,\ldots,k$, are denoted $\underline{x}_i$, and the matrix

$$X = \begin{pmatrix} \underline{x}_1' \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_k' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_{k1} & \cdots & x_{kp} \end{pmatrix},$$

is assumed to have rank p. The assumption that X has rank p simply assures that none of the column vectors of X is linearly redundant. The vector $(\beta_1,\ldots,\beta_p)'$ of regression coefficients is denoted $\underline{\beta}$. Thus the standardized mean difference for the $\underline{i}^{th}$ experiment is therefore $\delta_i = \underline{x}_i'\underline{\beta} = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p$. Denoting the vector of effect sizes by $\underline{\delta}$, i.e., $\underline{\delta}' = (\delta_1,\ldots,\delta_k)$ we can write the model for the effect sizes as

$$\underline{\delta} = X\underline{\beta}. \qquad (28)$$

Denote the vector of effect size estimates by $\underline{g} = (g_1, \ldots, g_k)'$.

Estimation of $\underline{\beta}$.

A model for the estimator $\underline{g}$ could be rewritten using $\underline{g}$, X, $\underline{\beta}$, and a residual vector $\underline{\eta}$ as

$$\underline{g} = X\underline{\beta} + \underline{\eta},$$

where $\underline{\eta}$ has the same distribution as $(\underline{g} - \underline{\delta})$, i.e.,

$$\underline{\eta} \sim N(0, \Sigma),$$

where $\Sigma = \mathrm{diag}(\sigma_1^2(\delta_1), \ldots, \sigma_k^2(\delta_k))$ and $\sigma_i^2(\delta_i)$ is given by (8). If the values of $\sigma_i^2(\delta_i)$ were known, we could use generalized least squares to obtain an estimator of $\underline{\beta}$. Unfortunately $\Sigma$ depends on $\underline{\delta}$ which is unknown. However, it is still possible to obtain estimates of $\underline{\beta}$ by using an estimated covariance matrix. Hedges (1982c) showed that the resulting estimator can be easily computed and has the same asymptotic distribution as the maximum likelihood estimator of $\underline{\beta}$. Therefore the alternative estimator is consistent and asymptotically efficient. This alternative estimator is also much easier to compute than is the maximum likelihood estimator.

Define the matrix V(g) as

$$V(g) = \mathrm{diag}(\sigma_1^2(g_1), \ldots, \sigma_k^2(g_k))$$

where $\sigma_i^2(g_i)$ is given by (8). An estimator $\hat{\underline{\beta}}$ of $\underline{\beta}$ under model (28) is given by

$$\hat{\underline{\beta}} = (X'V^{-1}(\underline{g})X)^{-1}X'V^{-1}(\underline{g})\underline{g}. \qquad (29)$$

The large sample approximation to the distribution of $\hat{\underline{\beta}}$ is given by

$$\hat{\underline{\beta}} \sim N_p(\underline{\beta}, \Sigma), \qquad (30)$$

where $\Sigma^{-1} = (\sigma^{st})$ s, t = 1,...,p and $\sigma^{st}$ is estimated by

$$\sigma^{st} = \sum_{i=1}^{k} \frac{x_{is}x_{it}}{\sigma_i^2(g_i)} \; ,$$

and $\sigma_i^2(g_i)$ is given by (8). Alternatively, $\Sigma = (X'V^{-1}(\underline{g})X)^{-1}$. There is also an iterated estimator of $\underline{\beta}$ that is analogous to (14), but the iterated version of $\hat{\underline{\beta}}$ rarely differs appreciably from (2) if the model is correctly specified.

The large sample approximation to distribution of $\hat{\underline{\beta}}$ can be used to provide approximate confidence intervals for the components of $\underline{\beta}$. That is, if $(X'V^{-1}(g)X)^{-1} = (V_{st})$, and $\hat{\underline{\beta}} = (\hat{\beta}_1, \ldots \hat{\beta}_p)'$, then a 100(1 - α) per cent confidence interval for $\beta_s$ is given by

$$\hat{\beta}_s - z_{\alpha/2}\sqrt{V_{ss}} \leq \beta_s \leq \hat{\beta}_s + z_{\alpha/2}\sqrt{V_{ss}} \, ,$$

where $z_{\alpha/2}$ is the 100(1 - α) per cent critical value of the normal distribution. The usual theory for the normal distribution can be used in conjunction with the Bonferroni inequality if simultaneous confidence intervals are desired.

Sometimes it is useful to test the hypothesis that $\underline{\beta} = 0$, that is, that all the components of $\underline{\beta}$ are simultaneously zero. The following statistics provide the basis for conducting these tests. The hypothesis that $\underline{\beta} = 0$ can be tested using the statistic

$$H_1 = \underline{\hat{\beta}}'X'V^{-1}(\underline{g})\underline{g}. \qquad (31)$$

If $\underline{\beta} = 0$, $H_1$ has an asymptotic chi-square distribution given by

$$H_1 \sim \chi^2_p \text{ .}$$

The test that $\underline{\beta} = 0$ at the significance level $\alpha$ therefore consists of comparing the obtained value of $H_1$ to the $100(1 - \alpha)$ per cent critical value of the chi-square distribution with p degrees of freedom. If the value of $H_1$ exceeds the critical value, the hypothesis that $\underline{\beta} = 0$ is rejected. Note that the statistic $H_1$ is analogous to the weighted sum of squares due to the regression in weighted least squares. Therefore the test that $\underline{\beta} = 0$ corresponds to a test that the weighted sum of squares due to the regression is greater than would be expected if $\underline{\beta} = 0$.

Testing Model Specification

We will argue subsequently that tests of model specification are an important step in the analysis of effect size data. The test of model specification is a way of determining whether the observed effect size estimates are reasonably consistent with the model used in the data analysis. If the number k of effect size estimates exceeds the number of predictors p, then a natural test of model specification is given by the next theorem. If k > p, the specification of the regression model can be tested using the statistic

$$H_2 = \underline{g}'V^{-1}(\underline{g})\underline{g} - H_1. \qquad (32)$$

When $\underline{\delta} = X\underline{\beta}$, i.e., the model is correctly specified, $H_2$ has an asymptotic chi-square distribution given by

$$H_2 \sim \chi^2_{k-p} \ .$$

The test for model specification at a significance level $\alpha$ therefore consists of comparing the obtained value of $H_2$ with the $100(1 - \alpha)$ per cent critical value of a chi-square distribution with (k - p) degrees of freedom. If the obtained value of $H_2$ exceeds the critical value, then model specification is rejected. Note that $H_2$ is analogous to the weighted residual sum of squares in weighted least squares. Thus the test for model specification is a test for greater residual variation than would be expected if $\underline{\delta} = X\underline{\beta}$. Such a test is not possible in the context of the usual normal theory because the means and variances of observations are independent. In the case of effect sizes, however, the sampling variance of $g_i$ given in (8) is completely determined by the mean of $g_i$ and the sample size. Therefore the "expected" residual variation is determined as a function of X and $\underline{\beta}$.

The test for model specification will often be used to demonstrate that the data (sample effect sizes) are reasonably consistent with the model used in data analysis. It is therefore important to have some understanding of the factors affecting the power of the test for model specification. Two factors that influence the power of the test are the number k of studies and the sample sizes ($n_i^E$ and $n_i^C$) of those studies. The latter factor (the sample sizes of the studies) is often the most significant. The reason is that the specification test statistic $H_2$ can be loosely described as a sum of squares of standardized residuals. The residual for the $\underline{i}^{th}$ study is "standardized" by the square root of the sampling variance of the $\underline{i}^{th}$ effect size estimate. When $n_i^E = n_i^C = n_i$, this sampling variance is approximately $2/n_i$. Therefore if the sample size $n_i$ in each group is large, even a small deviation from the model may result in a large contribution to the test statistic. Similarly, if the within-group sample sizes $n_i$ are small, even reasonably large deviations from the model may not yield a large "standardized residual" contribution to the test statistic. These arguments can be formalized into a rigorous development of power functions under so called local alternative hypotheses, but the formal arguments will not be given in this paper.

It is not necessarily true that large numbers of studies (large values of k) lead to rejection of model specification. The author has seen relatively simple models that fit well with over 100 studies, and many examples of well specified models for 40-80 effect sizes. If a particular model does not fit well, then diagnostic procedures are called for. Examination of residuals is often helpful. Such examinations may reveal patterns that suggest variables that should be added to the model. Alternatively, some studies may consistently yield effect size estimates that deviate greatly from the prediction of the model and therefore merit closer examination.

Computing Estimates and Test Statistics

The estimates and test statistics presented in this section can be easily calculated using any computer program package that manipulates matrices (such as SAS Proc Matrix). A simpler alternative to the computation of estimates and test statistics is the use of a computer program (such as SAS Proc GLM) that can perform weighted least squares analyses.

Weighted least squares involves estimation of linear model parameters by minimizing a weighted sum of squares of differences between observations and estimates. Given a design matrix X, a vector of observations of $\underline{Y}$, and a diagonal weight matrix W, the weighted least squares estimate of $\underline{\beta}$ in the model

$\underline{Y} = X\underline{\beta}$ is

$$\hat{\underline{\beta}}_W = (X'WX)^{-1}X'W\underline{Y}.$$

Note that the form of this estimator is the same as that of $\hat{\underline{\beta}}$. Thus $\hat{\underline{\beta}}$ is a special case of $\underline{\beta}_W$ where the weight matrix W is given by $V^{-1}(\underline{g})$. That is, the weight $i^{th}$ case is given by

$$w_i = \frac{2(n_i^E + n_i^C)n_i^E n_i^C}{2(n_i^E + n_i^C) + n_i^E n_i^C g_i}, \quad i = 1,\ldots,k,$$

and the weight matrix is $W = \text{diag}(w_1,\ldots,w_k)$.

The estimator $\hat{\underline{\beta}}$ is the weighted least squares estimator of $\underline{\beta}$ using design matrix X, data vector $\underline{g}$, and the weights $w_1,\ldots,w_k$ given above. The large sample covariance matrix of $\hat{\underline{\beta}}$ was given previously as $(X'V^{-1}(\underline{g})X)^{-1}$. This large sample covariance matrix is given by the weighted sum of squares and cross products matrix $(X'WX)^{-1}$ in the weighted least squares. If the computer program fits a "no-intercept" model, the test statistic $H_1$ for testing that $\underline{\beta} = 0$ is given by the weighted sum of squares due to the regression in the weighted least squares. A similar statistic for testing that all components of $\underline{\beta}$ except the intercept are simultaneously zero is given if the weighted least squares program fits an intercept. In the latter case the test statistic $H_2$ will be compared to the critical value of a chi-square distribution on k-p-1 degrees of freedom. The test statistic $H$ for testing model specification will always be the value of the weighted sum of squares about the regression line (the error sum of squares). Thus all of the statistics described in this section may be obtained from a single run of a standard packaged computer program.

## The Effects of Measurement Error on Effect Size

The standardized mean difference $\delta_i$ defined as in (1), is a measure of the magnitude of the treatment effect compared to the variability within the two groups of the experiment. The implicit assumption is that the variability within the experimental and control groups arises from stable difference between subjects (or more generally between expeimental units). If the response measure is not perfectly reliable, i.e., if errors of measurement are present, then measurement error also contributes to the within-group variability. Measurement error, therefore, alters the population value of the standardized mean difference. If the object is to estimate the value, $\delta$ , of the standardized mean difference when no errors of measurement are present, some procedure to correct for measurement error is necessary.

Consider the population value of the standardized mean difference in two cases, one in which the measurements are error-free and one in which errors of measurement are present. For simplicity of notation, the subscript i denoting the particular experiment is omitted in the exposition that follows, but the results apply to each experiment when properly indexed. If there are no errors of measurement, then denote the within-cell standard deviation by $\sigma_{\xi}^{2}$. Let $\delta$ denote the population value of the standardized mean difference when there are not errors of measurement. Then

$$\delta = (\mu^{E} - \mu^{C})/\sigma_{\xi} ,$$

where $\mu^{E}$ and $\mu^{C}$ are the population means of the experimental and control groups respectively. Note that the use of the symbol $\delta$ is consistent with the definition of $\delta$ used in the structural models (1).

In the second case, when errors of measurement are present, the population means $\mu^{E}$ and $\mu^{C}$ are unchanged but the within-group variance is larger. If $\delta'$ denotes the value of the standardized mean difference when errors of measurement are present, then

$$\delta' = (\mu^{E} - \mu^{C})/\sqrt{\sigma_{\xi}^{2} + \sigma_{\eta}^{2}} ,$$

where $\sigma_{\eta}^{2}$ is the variance due to errors of measurement. The relationship between $\delta$ and $\delta'$ can be expressed as

$$\delta' = \delta(\sigma_{\xi}/\sqrt{\sigma_{\xi}^{2} + \sigma_{\eta}^{2}}) = \delta\sqrt{\rho} ,$$

where $\rho$ is the reliability of the response measure.

Thus the population value of the standardized mean difference depends explicitly on the reliability ot the response measure. If the object is to estimate the value of $\delta$, the standardized mean difference with no errors of measurement, then estimation of $\delta'$ instead of $\delta$ can result in biased estimates. Since reliabilities cannot exceed one, the effect of measurement error is to reduce the magnitude of the parameter $\delta'$ compared with . In particular, errors of measurement cause the estimator g to estimate $\delta'$ instead of $\delta$, so that $E(g) = \delta' = \delta\sqrt{\rho}$. Hence errors of measurement result in underestimates of the parameter .

If the reliability $\rho$ is known, the bias can be removed by dividing g by $\sqrt{\rho}$. When we combine several estimates that use response scales with different reliabilities, each estimate can be corrected for measurement error separately. Statistical analyses can then be carried out using $g/\sqrt{\rho}$ in place of g and (g)/$\rho$ in place of (g).

## The Effects of Departures from Linear Equatability

In the statistical work described earlier we assumed that the tests used to measure outcomes in the different studies are linearly equatable. In practice this assumption may be only approximately true. Some tests may have unique factors in addition to the common factor shared among all tests. For example, some experiments may use an expensive standardized test to measure reading achievement, whereas other studies use locally developed tests that are correlated with the standardized test. If the locally developed tests have unique factors, they will not be perfectly valid measures of reading achievement as measured by the standardized test. This section reports some results of Hedges (1981) on the effect of invalidity of response measures on estimators of effect size.

One model for test invalidity assumes that a collection of tests share a common factor, but that some tests also have unique factors. If the population of test scores on a particular test are generated by a model which includes both the common factor (among all the tests) and a unique factor, then the test is partially invalid. To examine the effects of partial invalidity on effect size, Hedges (1981) derived the standardized mean difference $\delta''$ when tests had unique factors.

First consider the case where the treatment only affects the dependent variable via the common factor. Omitting the subscripts we see that the standardized mean difference is

$$\delta'' = \frac{\delta\sigma_{\xi}}{\sqrt{\sigma_{\xi}^{2} + \sigma_{\theta}^{2} + \sigma_{\eta}^{2}}} ,$$

where $\sigma_{\xi}^{2}$, $\sigma_{\theta}^{2}$, and $\sigma_{\eta}^{2}$ are the variances accounted for by the common factor, the unique factor, and measurement error, respectively. The validity coefficient $\rho^{V}$ of the test can be expressed as the within-group correlation of the test with the common factor, that is,

$$\rho^{V} = \frac{\sigma_{\xi}}{\sqrt{\sigma_{\xi}^{2} + \sigma_{\theta}^{2} + \sigma_{\eta}^{2}}} .$$

Therefore the population value of the standardized mean difference $\delta''$ can be expressed as

$$\delta'' = \delta\rho^V.$$

It seems unlikely that the population correlation of an invalid test with the common factor among a series of tests would be known. If X is a test that is not perfectly reliable, shares the Y common factor, but has no unique factor, then the correlation $\rho^V$ can be obtained from $\rho_{XY}$ by the familiar disattenuation formula (see, e.g., Lord and Novick, 1968):

$$\rho^V = \rho_{XY}/\sqrt{\rho} ,$$

where $\rho$ is the reliability of the test X. Thus the population standardized mean difference $\delta''$ can be written in terms of a correlation with a valid but unreliable test X and the reliability $\rho$ of X, namely,

$$\delta'' = \delta\rho_{XY}/\sqrt{\rho} .$$

Since $\rho_{XY} \leq \sqrt{\rho}$, it follows that $\delta'' \leq \delta$ . This means that invalidity always reduces the standardized mean difference when treatment affects only the common factor among the response measures. In this case, estimates of effect size may be corrected by substituting $g\sqrt{\rho}/\rho_{XY}$ for g.

When the treatment affects the test Y through both common and unique factors, invalidity of the test may either increase or decrease the standardized mean difference. Hence no simple characterization of the effect of invalidity on estimates of $\delta$ obtained from the estimators $g_i$ is possible. In this case the standardized mean difference is

$$\delta''' = \frac{\delta\sigma_\xi + \zeta}{\sqrt{\sigma_\xi^2 + \sigma_\theta^2 + \sigma_\eta^2}} ,$$

where $\zeta$ is the effect of the treatment on the unique factor; and $\sigma_\xi^2$, $\sigma_\theta^2$, and $\sigma_\eta^2$ are the variances due to the common factor, unique factor, and measurement error respectively. If $\zeta$, the treatment effect via the unique factor is large enough, namely

$$\zeta > (\sqrt{\sigma_\xi^2 + \sigma_\theta^2 + \sigma_\eta^2} - \sigma_\xi)\delta = \zeta_c ,$$

then $\delta''' > \delta$. If $\zeta < \zeta_c$, then $\delta''' < \delta$.

## Statistical Analysis When Correlations or Proportions are the Index of Effect Magnitude

In some cases, the effect size will not be a suitable index of effect magnitude for the studies that the reviewer wishes to integrate. For example, the studies may investigate the relationship between two continuous variables or they may study the proportion of subjects reaching a criterion in different groups. In the first example, Glass (1978) suggested the use of the correlation coefficient as an index of effect magnitude. In the second example, the difference between the proportions of subjects reaching the criterion appears to be a natural index of effect magnitude. Statistical procedures for combining correlation coefficients and differences between proportions were developed by Hedges and Olkin (1983). These procedures are analogous to the methods already presented for the analysis of effect sizes. One difference is that variance stabilizing transformations for correlations and proportions simplify the statistical methods.

## Statistical Analysis for Correlations as Effect Magnitudes

Suppose that k independent studies with sample sizes $n_1,\ldots,n_k$ yield k independent sample correlation coefficients $r_1,\ldots,r_k$. If $\rho_1,\ldots,\rho_k$ are the population correlations, the first problem is to decide if the sample correlations could reasonably have been drawn from populations with the same underlying population correlation. A test for the homogeneity of correlations is needed to determine if all the studies share a common population correlation. Statistical analyses are simplified if the correlations are transformed by Fisher's z-transformation. Let

$$z_i = \frac{1}{2}\log\left(\frac{1+r_i}{1-r_i}\right), \quad i = 1,\ldots,k.$$

A test for homogeneity of $\rho_1, \ldots, \rho_k$ uses the test statistic

$$H = \Sigma_{i=1}^{k} (n_i - 3)(z_i - z.)^2, \qquad (33)$$

where z. is the weighted average of the $z_i$ given by

$$z. = \frac{\Sigma_{i=1}^{k} (n_i - 3)z_i}{\Sigma_{i=1}^{k} (n_i - 3)} . \qquad (34)$$

If $\rho_1 = \rho_2 = \ldots = \rho_k$, and all the $n_i$ are moderately large, then H given in (33) is distributed approximately as a chi-square on (k - 1) degrees of freedom. Thus the test for homogeneity of the correlations consists of computing H and comparing the obtained value to the 100(1 - $\alpha$) per cent critical value of the chi-square distribution on (k - 1) degrees of freedom. If the obtained value of H exceeds the critical value we reject the hypothesis of homogeneity at the 100$\alpha$ per cent level. A computational formula for H is

$$H = \Sigma_{i=1}^{k} (n_i - 3)z_i^2 - \frac{\left( \Sigma_{i=1}^{k} (n_i - 3)z_i \right)^2}{\Sigma_{i=1}^{k} (n_i - 3)} .$$

If the correlations are homogeneous, then the natural estimate of the z-transform of the common correlation is z.

given in (34). This estimate may be converted into an estimate of $\rho$ by finding the value of $\rho$ that yields $z.$ as its z-transform, i.e.,

$$\hat{\rho} = \tanh(z.) = \frac{e^{2z.} - 1}{e^{2z.} + 1} . \qquad (35)$$

In large samples, z. has a normal distribution given by

$$z. \sim N(\zeta, \sigma_.^2),$$

where $\zeta$ is the z-transform of the common correlation $\rho$ and

$$\sigma_.^2 = \frac{1}{\sum_{i=1}^{k} (n_i - 3)} . \qquad (36)$$

This normal distribution can be used to obtain a large sample confidence interval for $\rho$. A 100(1 $-\alpha$) per cent confidence interval for $\zeta$ is given by

$$\zeta_1 = z. - z_{\alpha/2}\sigma. \leq \zeta \leq z. + z_{\alpha/2}\sigma. = \zeta_2,$$

where $\sigma_.^2$ is given in (36) and $z_{\alpha/2}$ is the 100(1 - $\alpha$) per cent critical value of the standard normal table. The 100(1 $-\alpha$ ) per cent confidence interval for $\rho$ is given by

$$\rho_1 = \tanh(\zeta_1) \leq \rho \leq \tanh(\zeta_2) = \rho_2.$$

where tanh(x) is given by (35).

If a priori knowledge or the formal test of homogeneity suggests that the correlations are not homogeneous, then the investigator may wish to determine whether various characteristics of the studies are related to the correlation.

Suppose that $\zeta_i$, the $z$-transform of the correlation in the $\underline{i}^{th}$ study depends on a vector of p study characteristics $(x_{i1},\ldots,x_{ip})'$ where $p\leq k$. The vectors $(x_{i1},\ldots,x_{ip})'$, $i = 1,\ldots,k$, are denoted $\underline{x}_i$ and define the design matrix X as

$$X = \begin{pmatrix} \underline{x}_1' \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_k' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_{k1} & \cdots & x_{kp} \end{pmatrix},$$

where X is assumed to have rank p. Define the vectors

$\underline{\zeta} = (\zeta_1,\ldots,\zeta_k)'$; $\underline{z} = (z_1,\ldots,z_k)'$, and $\underline{\beta} = (\beta_1,\ldots,\beta_p)'$. Then the model becomes $\zeta_i = x_{i1}\beta_1+\ldots+x_{ip}\beta_p$, $i = 1,\ldots,k$, or alternatively

$$\underline{\zeta} = X\underline{\beta} .$$

Hedges and Olkin (1983) showed that a natural estimator of $\underline{\beta}$ is

$$\hat{\underline{\beta}} = (X'VX)^{-1}X'V\underline{z}, \qquad (37)$$

where $V = \mathrm{diag}(n_1 - 3,\ldots,n_k - 3)$. When all the $n_i$ are reasonably large, $\hat{\underline{\beta}}$ has a p-variate normal distribution given by

$$\hat{\beta} \sim N_p(\underline{\beta}, (X'VX)^{-1}). \qquad (38)$$

The large sample distribution of $\hat{\underline{\beta}}$ can be used to obtain confidence intervals for $\beta_1, \ldots, \beta_p$. For example, a (nonsimultaneous) $100(1 - \alpha)$ per cent confidence interval for $\beta_j$ is given by

$$\hat{\beta}_j - z_{\alpha/2} v_{jj} \leq \beta_j \leq \hat{\beta}_j + z_{\alpha/2} v_{jj},$$

where $v_{jj}$ is the $\underline{j}^{th}$ diagonal element of $(X'VX)^{-1}$ and $z_{\alpha/2}$ is the $100(1 - \alpha)$ per cent critical value obtained from the standard normal distribution.

A simultaneous test that $\underline{\beta} = 0$ uses the test statistic

$$H_1 = \hat{\underline{\beta}}'(X'VX)\hat{\underline{\beta}}, \qquad (39)$$

which has a chi-square distribution when $\beta = 0$ given by

$$H_1 \sim \chi^2_p .$$

Thus the test that $\underline{\beta} = 0$ at a significance level $\alpha$ consists of comparing the obtained value of $H_1$ with the $100(1 - \alpha)$ per cent critical value of chi-square with p degrees of freedom. If $H_1$ exceeds the critical value we reject the hypothesis that $\underline{\beta} = 0$.

If $k > p$, a test of the specification (goodness of fit) of the regression model uses the test statistic

$$H_2 = \underline{z}'V\underline{z} - H_1 . \qquad (40)$$

When the model is correctly specified, the statistic $H_2$ has a chi-square distribution given by

$$H_2 \sim \chi^2_{k-p} .$$

Thus the test of model specification at a significance level $\alpha$ consists of comparing the obtained value of $H_2$ with the

$100(1 - \alpha)$ per cent critical value of the chi-square distribution on (k - p) degrees of freedom. If $H_2$ exceeds the critical value we reject the specification of the regression model. Rejection of model specification suggests that the model does not account for all of the variability of the correlations. In this case it may be desirable to think about additional explanatory variables or to examine residuals and look for unusual $z_i$ values to determine why the model does not fit well. It should be noted that the test for model specification can be very sensitive when sample sizes are large, so even minor deviations from the model can result in rejection of model specification.

## Statistical Analysis for Differences in Proportions

Suppose that k independent studies each compare the proportion of subjects achieving a criterion in an experimental and a control group. Let $p_i^E$ and $p_i^C$ denote the sample proportions of subjects reaching a criterion in the experimental and control groups respectively of the $\underline{i}^{th}$ study and let $\pi_i^E$ and $\pi_i^C$ represent the corresponding proportions in the population. The obvious index of the magnitude of the treatment effect in the $\underline{i}^{th}$ study is the difference between the experimental and control group proportions, $\pi_i^E - \pi_i^C$. Statistical analyses are simplified, however, if a slightly different index of effect magnitude is used. Define the population and sample indices of effect magnitude as

$$[illegible]_i = \sin^{-1}(\sqrt{\pi_i^E}) - \sin^{-1}(\sqrt{\pi_i^C}), \quad i = 1, \ldots, k, \qquad (41)$$

and

$$w_i = \sin^{-1}(\sqrt{p_i^E}) - \sin^{-1}(\sqrt{p_i^C}), \; i = 1,\ldots,k. \qquad (42)$$

A test of the hypothesis that $\omega_1 = \omega_2 = \ldots = \omega_k$ uses the test statistic

$$H = \sum_{i=1}^{k} 2\tilde{n}_i (w_i - w.)^2 \qquad (43)$$

where $\tilde{n}_i = n_i^E n_i^C/(n_i^E + n_i^C)$; $n_i^E$ and $n_i^C$ are the experimental and control group sample sizes in the $\underline{i}^{th}$ study, and w. is the weighted average of the $w_i$ given by

$$w. = \frac{\sum_{i=1}^{k} \tilde{n}_i w_i}{\sum_{i=1}^{k} \tilde{n}_i} . \qquad (44)$$

When $\omega_1 = \omega_2 = \ldots = \omega_k$ .and the sample sizes are all reasonably large, H given in (43) has a chi-square distribution on (k −1) degrees of freedom. Thus the test of homogeneity of the $\omega$'s at significance level $\alpha$ consists of comparing the obtained value of H with the $100(1 - \alpha)$ per cent critical value of the chi-square distribution with (k − 1) degrees of freedom. Values of H that are larger than the critical value result in rejection of the hypothesis that all studies share a common effect magnitude.

If the $w_i$ values could reasonably have come from populations with the same value of $\omega$, then the natural estimate of $\omega$ is $w.$ given in (44). In large samples the distribution of $w.$ is given by

$$w. \sim N(\omega,\sigma.^2),$$

where

$$\sigma.^2 = \frac{1}{\sum_{i=1}^{k} 2\tilde{n}_i} . \qquad (45)$$

This large sample distribution can be used to obtain an approximate confidence interval for $\omega$. A $100(1-\alpha)$ per cent confidence interval for $\omega$ is

$$w. - z_{\alpha/2}\sigma. \le \omega \le w. + z_{\alpha/2}\sigma.,$$

where $\sigma.^2$ is given in (45) and $z_{\alpha/2}$ is obtained from tables of the standard normal distribution.

If a priori knowledge or a formal hypothesis test lead the investigator to believe that the $\omega$'s are not homogeneous, then the investigator may wish to determine whether various characteristics of studies are related to the index $\omega_i$ of effect magnitude. Suppose that $\omega_i$ depends on a vector of p study characteristics $\underline{x}'_i = (x_{i1},\ldots,x_{ip})'$, where $p \le k$. The vectors $\underline{x}_1,\ldots,\underline{x}_k$ define the design matrix

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_k' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{k1} & \cdots & x_{kp} \end{pmatrix}$$

where X is assumed to have rank p. Define the vectors $\underline{\omega} = (\omega_1,\ldots,\omega_k)'$ and $\underline{w} = (w_1,\ldots,w_k)'$ of parameters and estimates and let $\underline{\beta} = (\beta_1,\ldots,\beta_p)'$ be the vector of regression coefficients. The linear model is $\omega_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p$, i = 1,...,k, or alternatively $\underline{\omega} = X\underline{\beta}$.

Hedges and Olkin (1983) showed that the natural estimator of $\underline{\beta}$ is

$$\underline{\hat{\beta}} = (X'VX)^{-1}X'V\underline{w}, \qquad (46)$$

where $V = 2\mathrm{diag}(\tilde{n}_1,\ldots,\tilde{n}_k)$. When all of the $\tilde{n}_i$ are reasonably large, $\underline{\hat{\beta}}$ has a p-variate normal distribution given by

$$\hat{\beta} \sim N_p(\underline{\beta},(X'VX)^{-1}).$$

This large sample distribution of $\underline{\hat{\beta}}$ can be used to obtain confidence intervals for $\hat{\beta}_j$ just as in the case of the analysis of correlations. The test that $\underline{\beta} = 0$ and the test of model specification based on the statistics

$$H_1 = \underline{\hat{\beta}}'(X'VX)\underline{\hat{\beta}}$$

and

$$H_2 = \underline{w}'V\underline{w} - H_1$$

are identical to the analogous tests in the analysis based on correlations.

## CONCLUSIONS

There has been some vehement criticism of Glass's meta-analysis. Some critics have argued that meta-analysis may lead to oversimplified conclusions about the effect of a treatment because it condenses the results of a series of studies into a few parameter estimates. For example, Presby (1978) argued that even when studies are grouped according to variations in the treatment, reviewers might reasonably disagree on the appropriate groupings. Grouping studies into overly broad categories and calculating a mean effect size for each category might serve to wash out real variations among treatments in the categories. Thus it would appear that variations in treatment were unrelated because the mean effect sizes for the categories did not differ. An obvious extension of this argument is that reviewers might reasonably disagree on explanatory variables that could be related to effect sizes. Hence failure to find variables that are systematically related to effect size does not imply that the effect sizes are consistent across studies. It may only imply that the reviewer has examined the wrong explanatory variables.

A related criticism is that the studies in a collection may give fundamentally different answers (e.g., have different population effect sizes) perhaps because of the artifacts of a multitude of design flaws (see e.g., Eysenck, 1978). Any analysis of the effect sizes is therefore an analysis of estimates influenced by a variety of factors other than the true magnitude of the effect of the treatment. Thus meta-analyses may be another case of "garbage in--garbage out." The argument underlying this criticism is that flaws in studies may influence effect sizes.

The statistical methods presented in this monograph provide a mechanism for responding to criticisms mentioned previously. In the simplest case the reviewer summarizes the results of a series of studies by the average effect size estimate. Is this an oversimplification of the results of the studies? The test of homogeneity of effect size provides a method of empirically testing whether the variation in effect size estimates is greater than would be expected by chance alone. If the hypothesis of homogeneity is not rejected, the reviewer is in a strong position vis-a-vis the argument that studies exhibit real variability which is obscured by course grouping. If the model of a single population effet size fits the data adequately, then a desire for parsimony suggests this model should be considered seriously.

Failure to reject the homogeneity of effect sizes from a series of studies does not necessarily disarm the criticism that the results of the studies are artifacts of design flaws. For example, if a series of studies all share the <u>same</u> flaw, consistent results across the series of studies may be an artifact of just that flaw. That is, the design flaw in all of the studies may act to make the effect sizes in the studies consistent with one another and consistently wrong as an estimate of the treatment effect. On the other hand, the studies may not all have the same flaws. If a variety of different studies, with different design flaws all yield consistent results it may be implausible to explain the consistency of the results of a series of studies as a conspiracy of different artifacts all yielding the same bias. Thus the reviewer who finds consistency in research results and who knows the limitations of the individual studies may be in a strong position against the "garbage in--garbage out" argument. I emphasize that careful examination of the individual research studies and some scrutiny of the attendant design problems is essential. Without such analysis of the studies, a single source of bias is a very real and plausible rival explanation for empirical consistency of research results.

When a reviewer explains the effect sizes from a series of studies via a model involving explanatory variables (e.g., the effect size varies according to grade level), tests of model specification play a role analogous to that of the test of homogeneity. It is difficult to argue that additional variables are needed to explain the variation in effect sizes if the specification test suggests that additional variables are not needed.

Evidence that the model is correctly specified does not necessarily mean that the artifacts of design flaws may be ignored. Il all studies share a common design flaw then the results of all of the studies may be biased to an unknown extent. If design flaws are correlated with explanatory variables, then the effects of those design flaws are confounded with the effects of the explanatory variable . It may be difficult or impossible to determine the real source of the effect. However, if several design flaws occur in different studies, if few of the flaws occur in more than a few studies, and if simple models appear to be correctly specified, then it seems implausible that inferences drawn about the effect sizes are artifacts of biases due to design flaws.

Methods for the quantitative synthesis of research (meta-analysis) have received a great deal of attention. These methods are promising improvements in the methodology for research reviewing. Some authors (Jackson, 1980) have argued that research reviewing in social sciences has not been subject to the kind of rigorous professional standards of methodology that have been imposed on the conduct of primary research. One positive aspect of the interest in meta-analysis has been the movement toward the use of rigorous methods of sampling, measurement, and statistics in research reviewing (Cooper, 1982). As Glass (1978) pointed out, research reviewing is a difficult task that demands as much creativity and insight as the conduct of primary research. The application of rigorous methods in research reviews may be a step toward a time when reviewing is accorded equal status with the generation of primary research.

Statistical techniques for research reviews are not well developed. The present survey of methodology is a beginning. New methods are clearly needed for the exploratory analysis of effect size data. Glass (1978) argued that it is often useful to estimate effect sizes from other statistics or from data other than means and standard deviations. Methods for obtaining such estimates were discussed in Glass (1978) and in Glass, McGaw, and Smith (1981). Yet very little is known about the properties of such estimates. The development of rigorous statistical theory for these estimators would be an important contribution.

Many research studies provide data on several measures of the same or related constructs. The several effect size estimates derived from different measures applied to the same individual are therefore correlated. In fact, the vector of effect size estimates can be shown to have a multivariate normal distribution in large samples. Moreover, the (large sample) correlation matrix of the effect sizes is the same as that of the original observations. Thus if a reading and a mathematics achievement test have a (population) correlation of .7, effect sizes derived from these two tests will also have a correlation of .7 in large samples. This result can be used in the study of covariation among effect sizes derived form the same sample, but relatively little work has been done. A conservative solution is to use the average of multiple effect sizes as the only estimate of effect size in statistical analyses. More work is definitely needed in the area of the multivariate analysis of effect sizes. A related issue is how to handle estimation of effect size from a series of correlated estimates. Glass (1978) suggested the use of Jackknife estimators in this case, which seems sensible. This problem of estimation of effect size from correlated estimates merits further investigation.

The possibility of influences of bias due to the use of statistical significance as a criterion in editorial decisions has been suggested (Sterling, 1959). Some preliminary work (Lane & Dunlap, 1978) suggests that the effects of publication bias can be severe. We have little evidence about how seriously such biases affect quantitative research syntheses. Some empirical evidence on this question is discussed in Glass, McGaw, and Smith (1981). There are very few statistical methods for dealing with the effects of these biases. Future meta-analyses will, no doubt, reveal new methodological problems that also need attention.

## References

Brewer, J. K. On the power of statistical tests used in the American Educational Research Journal. *American Educational Research Journal,* 1972, *9*, 391-401.

Campbell, D. T. Definitional versus multiple operationalism. *Et al.*, 1969, *2*, 14-17.

Chase, L. J., & Chase, R. B. A. A statisticıl power analysis of applied psychological research. *Journal of Applied Psychology,* 1976, *61*, 234-237.

Cochran, W. G. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society Supplement,* 1937, *4*, 102-118.

Cochran, G. W. The combination of estimates from different experiments. *Biometrics,* 1954, *10*, 101-129.

Cohen, J. The Statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 1962, *65*, 145-153.

Cohen, J. *Statistical power analysis for the behavioral sciences.* New York : Academic Press, 1979.

Cooper, H. M. Scientific guidelines for conducting integrative research reviews. *Review of Educational Research,* 1982, *52*, 291-302.

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

Gage, N. L. *The scientific basis of the art of teaching.* New York: Teachers College Press, 1978.

Giaconia, R. M., & Hedges, L. V. Identifying features of etiective open education. *Review of Educational Research,* 1982, *52*, (in press).

Glass, G. V Primary, secondary, and meta-analysis of research. *Educational Researcher,* 1976, *5*, 3-8.

Glass, G. V Integrating findings: The meta-analysis of research. In L. S. Schulman (Ed.) *Review of Research in Education, 5*, Itasca, Ill. : F. E. Peacock, 1978.

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

Glass, G. V & Smith, M. L. Meta-analysis of the relationship between class-size and achievement. *Educational Evaluation and Policy Studies,* 1979, *1*, 2-16.

Hedges, L. V. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics,* 1981, *6*, 107-128.

Hedges, L. V. Estimating effect size from a series of independent experiments. Psychological Bulletin, 1982, 92, 490-499. (a)

Hedges, L. V. Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 1982, 7, 119-137. (b)

Hedges, L. V. Fitting continuous models to effct size data. Journal of Educational Statistics, 1982, 7, (in press). (c)

Hedges, L. V., & Olkin, I. Vote counting methods in research synthesis. Psychological Bulletin, 1980, 88, 359-369.

Hedges, L. V., & Olkin, I. Regression models in research synthesis. American Statistician, 1983, 37, (in press).

Jackson, G. B. Methods for integrative reviews. Review of Educational Research, 1980, 50, 438-460.

Katzer, J., & Sodt, J. An analysis of the use of statistical testing in communications research. Journal of Communication, 1973, 23, 251-265.

Kraemer, H. C., & Andrews, G. A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 1982, 91, 404-412.

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

Lane, D. M., & Dunlap, W. P. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, 1978,31, 107-112.

Li , R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Education Review, 1971, 41, 429-471.

Lord, F. L., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Pascarella, E. T., Walberg, H. J., Junker, L. K., & Haertel, G. D. Continuing motivation in science for early and late adolescents. American Educational Research Journal, 1981, 18, 439-452.

Pearson, K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 1933, 25, 379-410.

pillemer, D. B., & Light, R. J. Synthesizing outcomes: How to use research evidence from many studies. *Harvard Education Review;* 1980, *50*, 176-195.

Presby, S. Overly broad categories obscure important differences. *American Psychologist,* 1978, *33*, 514-515.

Rosenthal, R. R., & Rubin, D. B. Comparing effect sizes of independent studies. *Psychological Bulletin,* 1982, *92*, 500-504.

Smith, M. L., & Glass, G. V Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 1977, *32*, 752-760.

Smith, M. L., & Glass, G. V Meta-analysis of class size and its relationship to attitudes and instruction. *American Educational Research Journal,* 1980, *17*, 419-433.

Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance--or visa versa. *Journal of the American Statistical Association,* 1959, *54*, 30-34.

Tippett, L. H. C. *The methods of statistics.* New York: Wiley, 1931.

Uguroglu, M. E., & Walberg, H. J. Motivation and achievement: A quantitative synthesis. *American Educational Research Journal,* 1979, *16*, 375-390.

Yates, F., & Cochran, W. G. The analysis of groups of experiments. *Journal of Agricultural Science*, 1938, *28*, 556-580.

ERIC/TM Report 83

# STATISTICAL METHODOLOGY IN META-ANALYSIS

BY

Larry V. Hedges

The University of Chicago

Many problems must be addressed by the reviewer who carries out a meta-analysis. These problems include identifying and obtaining appropriate studies, extracting estimates of effect size from the studies, coding or classifying studies, analyzing the data, and reporting the results of the data analysis. Earlier work by Glass, McGaw, and Smith describes methods for dealing with these problems.

However, since their book, there has been a great deal of interest in the development of systematic statistical theory for meta-analysis. The purpose of this monograph is to provide a unified treatment of rigorous statistical methods for meta-analysis. These methods provide a mechanism for responding to criticisms of meta-analysis, such as that meta-analysis may lead to oversimplified conclusions or be influenced by design flaws in the original research studies.

---

ORDER FORM

Please send_______copies of ERIC/TM Report 83, "Statistical Methodology in Meta-Analysis," at $7.00.

Name________________________________

Address_____________________________

____________________________Zip_________

Total enclosed $___________

Return this form to:

ERIC/TM
Educational Testing Service
Princeton, NJ 08541

## RECENT TITLES

## IN THE ERIC/TM REPORT SERIES

#83 - Statistical Methodology in Meta-Analysis, by Larry V. Hedges. 12/82
$7.00.

#82 - Microcomputers in Educational Research, by Craig W. Johnson. 12/82
$8.50.

#81 - A Bibliography to Accompany the Joint Committee's Standards on
Educational Evaluation, compiled by Barbara M. Wildemuth. 107p.
$8.50.

#80 - The Evaluation of College Remedial Programs, by Jeffrey K. Smith
and others. 12/81. $8.50.

#79 - An Introduction to Rasch's Measurement Model, by Jan-Eric Gustafsson.
12/81, $5.50.

#78 - How Attitudes Are Measured: A Review of Investigations of Professional,
Peer, and Parent Attitudes toward the Handicapped, by Marcia D. Horne.
12/80, $5.50.

#77 - The Reviewing Processes in Social Science Publications: A Review of
Research; by Susan E. Hensley, and Carnot E. Nelson. 12/80, $4.00.

#76 - Intelligence Testing, Education, and Chicanos: An Essay in Social
Inequality, by Adalberto Acquirre Jr. 12/80, $5.50.

#75 - Contract Grading, by Hugh Taylor. 12/80, $7.50.

#74 - Intelligence, Intelligence Testing and School Practices, by Richard
DeLisi. 12/80, $4.50.

#73 - Measuring Attitudes Toward Reading, by Ira Epstein. 12/80, $9.50.

#72 - Methods of Identifying Gifted Minority Students, by Ernest M. Bernal.
12/80, $4.50.

#71 - Sex Bias in Testing: An Annotated Bibliography, by Barbara Hunt. 12/79,
$5.00.

#70 - The Role of Measurement in the Process of Instruction, by Jeffrey K.
Smith. 12/79, $3.50.

#68 - The Educational Implications of Piaget's Theory and Assessment
Techniques, by Richard DeLisi. 11/79, $5.00.

#66 - Competency-Based Graduation Requirements: A Point of View, by Mary
Ann Bunda. 1978, $2.00.

#65 - The Practice of Evaluation, by Clare Rose and Glenn F. Nyre. 12/77,
$5.00.

#63 - Perspectives on Mastery Learning & Mastery Testing, by Jeffrey K.
Smith. 1977, $3.00.