DOCUMENT RESUME

ED 276 759 TM 860 710

AUTHOR Kulik, James A.; Kulik, Chen-Lin C.
TITLE Operative and Interpretatbie Effect Sizes in Meta-Analysis.
PUB DATE Apr 86
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; Error of Measurement; Estimation (Mathematics); *Mathematical Models; *Meta Analysis; Pretests Posttests; *Research Methodology; Social Science Research; Statistical Studies; Test Interpretation
IDENTIFIERS *Effect Size Estimator (Glass); *Glass Analysis Method

ABSTRACT

Statistical methodologists have sometimes criticized the use of conventional statistics in meta-analysis, and in recent years a number of them have advocated the use of a special new statistical methodology for research synthesis. An examination of recent books describing this methodology shows that it is seriously limited in its applicability to social science research findings. The new methodology produces interpretable meta-analytic results only in exceptional circumstances (e.g., when each study in a collection uses the same unblocked, posttest-only experimental design). The new statistical methodology for meta-analysis has produced uninterpretable results when applied to typical collections of social science studies with varied experimental designs. (Author)

# Operative and Interpretable Effect Sizes In Meta-analysis

James A. Kulik & Chen-Lin C. Kulik

The University of Michigan



A paper presentation at the annual meeting
Of the American Educational Research Association,
San Francisco, April 1986

## Abstract

Statistical methodologists have sometimes criticized the use of conventional statistics in meta-analysis, and in recent years a number of them have advocated the use of a special new statistical methodology for research synthesis. An examination of recent books describing this methodology shows that it is seriously limited in its applicability to social science research findings. The new methodology produces interpretable meta-analytic results only in exceptional circumstances (e.g., when each study in a collection uses the same unblocked, posttest-only experimental design). The new statistical methodology for meta-analysis has produced uninterpretable results when applied to typical collections of social science studies with varied experimental designs.

## Operative and Interpretable Effect Sizes in Meta-analysis

In a classic 1976 paper Glass defined meta-analysis as the application of statistical methods to results from a large collection of studies for the purpose of integrating the findings. The statistical methods that Glass used in meta-analysis were conventional ones, such as analysis of variance and multiple regression analysis. In meta-analysis, however, these statistics were applied not to raw observations, but rather to effect sizes, or standardized scores that represented the size of treatment effects in different studies on a common scale of standard deviation units. Hedges (1984) has recently commented on Glass's use of conventional statistics in research synthesis:

> Such use seemed at first to be an innocuous extension of statistical methods to a new situation. However, recent research has demonstrated that the use of such statistical procedures as analysis of variance and regression analysis cannot be justified for meta-analysis. Fortunately, some new statistical procedures have been designed specifically for meta-analysis (p. 25).

This new methodology for meta-analysis builds on statistical techniques originally developed by Cochran (1954) for testing the homogeneity of results in related experiments and for making composite estimates of population parameters from such results. Hedges (1981, 1982) first showed how Cochran's techniques could be applied to experimental results coded as effect sizes. Hunter, Schmidt, and Jackson (1982) and Rosenthal (1984) later advocated the use of formulas and tests very similar to those used by Hedges. Hedges (1984) has referred to the methods used by himself, Rosenthal, and Hunter and Schmidt as "modern statistical methods for meta-analysis." Bangert-Drowns (in press) has described the methods as "approximate data pooling."

Recent books describing these methods (Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; and Rosenthal, 1984) almost entirely ignore one of the central emphases in Glass's work: the estimation of effect sizes from studies with different experimental designs. Glass and his colleagues have argued that different procedures are needed for estimating effect sizes in simple experiments with unblocked posttest-only designs and in more complex experiments (Glass, McGaw, & Smith, 1981; McGaw & Glass, 1980). Glass and his colleagues use the following formula, for example, to estimate effect sizes from simple experiments that compare posttest means of independent groups:

$$\underline{d} = \frac{\underline{M}_{\underline{Y}}^{E} - \underline{M}_{\underline{Y}}^{C}}{\underline{S}_{\underline{Y}}}, \qquad (1)$$

where $\underline{d}$ is the estimator of effect size for a specific study, $\underline{M}_{\underline{Y}}^{E}$ and $\underline{M}_{\underline{Y}}^{C}$ are sample means of the experimental group (E) and control group (C) on the dependent variable $\underline{Y}$, and $\underline{S}_{\underline{Y}}$ is the sample standard deviation on $\underline{Y}$. They use different formulas for calculating effect sizes from more powerful experimental designs that control for irrelevant sources of variation in $\underline{Y}$: comparisons of matched groups, comparisons of gain scores, covariance analyses, multifactor analyses of variance, etc. (Glass et al., 1981, pp. 114-123).

The recent books on meta-analytic methodology all give Formula 1 as a basic formula for estimating effect sizes, but none of the books quotes even one of the additional formulas that Glass and his colleagues have used for calculating effect sizes from studies with more complex experimental designs. Hedges and Olkin (1985, p. 13), for example, have simply noted that such formulas are outside their domain of interest; they do not consider them a central issue in the statistics of meta-analysis. Rosenthal (1984, pp. 30-31) has referred to these formulas as "formulas for adjusting effect sizes" and has cautioned that those who calculate them should also report "unadjusted effect sizes" alongside the "adjusted" ones.

Neither Rosenthal nor other developers of the newer statistical methods for meta-analysis, however, have written much about the calculation of "unadjusted" effect sizes with powerful experimental designs. Recent books on meta-analysis focus almost exclusively on calculation of effect sizes with an unblocked, posttest-only design. The only design other than this one covered in detail in recent books is the comparison of pre-post gains of experimental and control groups. What recent methodologists have said about this design shows how different their approach is from Glass's.

Glass and his colleagues provided the following formula for calculating an effect size for this design:

$$d_G = \frac{M_G^E - M_G^C}{S_Y}, \qquad (2)$$

where $d_G$ is the effect size estimated from a comparison of gain scores and $M_G^E$ and $M_G^C$ are the average gains for the experimental and control groups (Glass et al., 1981, pp. 115-118). Note that the numerator of Formulas 1 and 2 are calculated differently. The numerator of Formula 2 is calculated from group gains rather than from group postscores. But formulas (1) and (2) do not differ in denominators. For both designs, the posttest standard deviation is used in standardizing the mean differences.

Rosenthal, on the other hand, has provided the following formula for use with gains analyses (1985, p. 21):

$$d'_G = \frac{M_G^E - M_G^C}{S_G}, \qquad (3)$$

where $d'_G$ is the effect size estimated by Rosenthal from a comparison of gain scores; $M_G^E$ and $M_G^C$ are defined as above; and $S_G$ is the standard deviation of the gain scores and is equal to $S_Y\sqrt{2(1 - r_{XY})}$, where $r_{XY}$ is the correlation between pretest ($X$) and posttest ($Y$) scores. Note that Rosenthal and Glass's formulas differ in standardization term. Rosenthal uses the standard deviation of gain scores in the denominator of his formula, whereas Glass uses the standard deviation of the posttest in his formula. Effect sizes calculated by Glass and Rosenthal from the same gains analysis would therefore be related by the following formula:

$$d'_G = \frac{d_G}{\sqrt{2(1 - r_{XY})}}. \qquad (4)$$

Because pretest-posttest $r$'s can be quite high in educational and psychological research, effect sizes calculated by Glass and

Rosenthal can differ by a large amount. For example, with a pretest-posttest correlation of .8, an effect size calculated from Formula 2 would be only 60% as large as an effect size calculated from Formula 3 for the same experiment.

Kraemer and Andrews (1982), who have also contributed to the development of new statistical methods for meta-analysis, have criticized Formula 3 on the grounds that the standardization term is wrong. They have pointed out that Formulas 1 and 3 do not estimate the same quantity. But what is remarkable about Kraemer and Andrews' discussion is their tacit assumption that meta-analysts use Formula 3 to calculate effect sizes from comparisons of gains. Long before Kraemer and Andrews published their article, Glass and his colleagues had advised meta-analysts not to use Formula 3 in calculating effect sizes for comparisons of gains but to use Formula 2 instead. Glass's advice has been either overlooked, ignored, or misunderstood by recent contributors to meta-analytic methodology.

The purpose of this article is to review issues involved in estimating effect sizes from studies that use different experimental designs. It covers estimation of effect sizes for both simple posttest-only designs and complex designs that remove sources of irrelevant variation from the posttest. This article also presents formulas for calculating the standard error of effect sizes estimated under different conditions. Finally, the article shows that average effect sizes and their standard errors are often miscalculated in meta-analyses that use the newer statistical methods for research synthesis.

## Operative and Interpretable Effect Sizes

The notion of measuring effect sizes was a familiar one to many social scientists long before Glass used indices of effect sizes as a key tool in meta-analysis. Cohen's (1977) classic book on power analysis in the social sciences made extensive use of effect sizes in estimating the power of statistical tests and in determining sample sizes needed to achieve tests of a given power. Cohen's book on power analysis also introduced a critical distinction between two types of effect sizes: <u>interpretable effect sizes</u> and <u>operative effect sizes</u>.

When calculated from means, interpretable effect sizes are determined by dividing a treatment effect expressed in raw units (<u>Y</u>-units) by the standard deviation of <u>Y</u>. Cohen used the symbol <u>d</u> to stand for the interpretable effect size for the posttest-only, independent-group design. He added primes and subscripts to the symbol <u>d</u> to denote interpretable effect sizes calculated for other

experimental designs. For example, Cohen used the symbol $\underline{d}'_4$ for the interpretable effect size calculated for one sample of $\underline{n}$ differences between paired observations (1977, p. 49):

$$\underline{d}'_4 = \frac{\underline{M}^{E}_{\underline{Y}} - \underline{M}^{E}_{\underline{X}}}{\underline{S}_{\underline{Y}}}, \qquad (5)$$

where $\underline{M}^{E}_{\underline{X}}$ is the mean of the experimental group on a pretest and $\underline{M}^{E}_{\underline{Y}}$ is the mean on the posttest. Cohen used the symbol $\underline{d}'_3$ for the interpretable effect size calculated for a study in which the mean of one experimental group is compared to a theoretical population mean (p. 46). Cohen pointed out, however, that all such interpretable effect sizes are conceptually equivalent and can be interpreted on a common scale. This is because the standardizing unit for interpretable effect sizes is always the standard deviation of $\underline{Y}$ ($\underline{S}_{\underline{Y}}$).

Operative effect sizes are an entirely different matter. Operative effect sizes are calculated by dividing a treatment effect expressed in $\underline{Y}$-units by either the standard deviation of $\underline{Y}$ or by a standard deviation from which major sources of variation have been removed by one or another adjusting mechanism designed to increase power--covariance, regression, or blocking. Operative effect size are identical to interpretable effect sizes only in experiments that do not remove sources of irrelevant variation from the dependent variable, e.g., Campbell and Stanley's (1963) Type 6 experiments. For other experimental designs, operative effect sizes are calculated with special formulas. The operative effect size $\underline{d}$ for paired observations for one sample would be estimated from (Cohen, 1977, p. 63):

$$\underline{d} = \frac{\underline{M}^{E}_{\underline{Y}} - \underline{M}^{E}_{\underline{X}}}{\underline{S}_{\underline{Y}}\sqrt{1 - \underline{r}_{\underline{XY}}}}. \qquad (6)$$

Cohen used the symbol $\underline{d}$, without subscripts or primes, to represent operative effect sizes calculated for a variety of experimental designs. Although denoted by a common symbol, operative effect sizes calculated for different experimental designs are not conceptually equivalent because different

standardizing units are used in calculating them. Operative effect sizes cannot therefore be interpreted in a single way. Operative effect sizes are useful, however, because they can be employed directly to find values of power in power tables.

A critical point to grasp is this: Meta-analysis must be based on interpretable effect sizes, not operative effect sizes, if it is to produce interpretable results. Operative effect sizes have an undesirable property that makes them inappropriate to use in meta-analysis. With a given raw-score unit, they vary not only as a function of size of the raw-score treatment effect but also as a function of the experimental design used to investigate the effect. Two investigators studying the same phenomenon who find identical treatment effects (when effects are expressed in raw-score units) would report very different operative effect sizes if they used different research designs to study this treatment effect.

Suppose, for example, that Investigators A and B are studying the effect of a diet-supplement program, and each investigator finds that the program has the same effect on the dependent variable: It adds an average of 10 pounds to the weight of experimental subjects. Suppose further that Investigator A has used a weak experimental design to estimate the effect and test its significance (e.g., a posttest-only design for independent groups), whereas investigator B has used a more powerful design (e.g., analysis of covariance with a covariate that correlates .9 with the dependent variable). Even though the raw treatment effects estimated by the two investigators are of the same magnitude, the operative effect size calculated by investigator B will be nearly twice the size of the operative effect size calculated by investigator A because the standardization term employed by investigator A is the raw standard deviation, whereas the standardization term employed by Investigator B is a standard deviation of residuals from which important sources of variation have been removed. The interpretable effect sizes calculated by the two investigators, however, will be the same--as they should be for two raw treatment effects of the same magnitude.

It is useful to keep this distinction between operative and interpretable effect sizes in mind when examining Formulas 2 and 3, the formulas that Glass and Rosenthal use for calculating effect sizes from comparisons of gains. It should be clear that the formula used by Glass is a formula for an interpretable effect size; Rosenthal's formula is for an operative effect size. It should also be clear that Glass's formula is the appropriate one to use in meta-analysis, and Rosenthal's formula is inappropriate for use in research synthesis.

Although Glass and his colleagues have not mentioned the distinction between operative and interpretable effect sizes in their writings on meta-analysis, their work suggests that they were aware of the importance of using interpretable effect sizes, rather than operative ones, in research syntheses (e.g., Glass et al., 1981). The various formulas for effect size that they have given are all formulas for interpretable effect sizes. Rosenthal (1984), on the other hand, appears to believe that operative effect sizes are the appropriate ones to calculate in research synthesis. Although Kraemer and Andrews (1982) appear to recognize the inappropriateness of operative effect sizes, they appear to believe that such effect sizes are the ones that meta-analysts actually calculate.

## Standard Error of an Effect Size

One of the major contributions of Hedges to the methodology for meta-analysis is the derivation of a formula for the standard error of effect sizes. Hedges (1982) derived this formula for standard error from a set of explicit assumptions about the data being analyzed in research syntheses. Careful examination of these assumptions will reveal, we believe, that they seldom are met by meta-analytic data sets.

Data in the model that underlies Hedges' formulas for analysis of effect sizes come from a series of $k$ independent studies, each of which compares an experimental group (E) with a control group (C). Hedges lets $Y^E_{ij}$ and $Y^C_{ij}$ stand for the $j$th scores in the $i$th experiment from the experimental and control groups, respectively. His model assumes that in a given study $i$, $Y^E_{ij}$ and $Y^C_{ij}$ are normally distributed with means $\mu^E$ and $\mu^C$ and common variance $\sigma$. The population effect size for study $i$ will then be given by the following (Hedges, 1982, p. 491):

$$\delta = \frac{\mu^E - \mu^C}{\sigma} . \qquad (7)$$

Hedges has examined the properties of Glass's estimator $d$ (Equation 1) of this population effect size. He has shown that in small samples Glass's $d$ is a biased estimator of the population parameter. An unbiased estimator of $\delta$ can be approximated from the following:

$$\underline{d}^{U} = \left[1 - \frac{3}{4(\underline{n}^{E} + \underline{n}^{C} - 2) - 1}\right] \underline{d} \ , \qquad (8)$$

where $\underline{n}^{E}$ and $\underline{n}^{C}$ are the sample sizes of the experimental and control groups. The sampling distribution of this unbiased estimator is that of a noncentral $\underline{t}$ times a constant. If both $\underline{n}^{E}$ and $\underline{n}^{C}$ are large, the distribution of $\underline{d}^{U}$ or $\underline{d}$ is approximately normal with $\mu(\underline{d}) = \delta$ and

$$\sigma(\underline{d}) = \sqrt{\left[\frac{1}{\underline{n}^{E}} + \frac{1}{\underline{n}^{C}}\right] + \frac{\delta^{2}}{2(\underline{n}^{E} + \underline{n}^{C})}} \ , \qquad (9)$$

where $\delta$, the population effect size, is the noncentrality parameter (Hedges, 1982, p. 492). This is an important formula because it can be used to approximate the sampling error of effect sizes estimated under certain conditions.

It is important to emphasize, however, that the formula applies only to situations that meet the assumptions made by Hedges. These assumptions include (a) independence of experimental and control groups, and (b) assessment of results on a dependent variable from which no sources of irrelevant variation have been removed in the experimental design. The model fits a type of design that Campbell and Stanley (1963) call a Type 6 design, the posttest-only design for independent groups. It is a design from which valid inferences can be drawn about experimental treatments, but researchers who use it do not estimate treatment effects with great precision.

Many social science researchers prefer to conduct experiments that estimate treatment effects more precisely. Instead of posttest-only designs for independent groups, they use experimental designs and statistical techniques that remove sources of irrelevant variation from their dependent variables. They use multiple factor or matched-subject designs; they compare gain scores of experimental and control groups; or they include covariates in their statistical analyses. Such designs, rather than the posttest-only design for independent groups, dominate the literature of the social sciences. For example, none of the 14 studies included in a recent meta-analysis on coaching effects on the Scholastic Aptitude Test (SAT) used a posttest-only design for

independent groups (Kulik et al., 1984). All studies located for this meta-analysis compared gains or covariance-adjusted postscores of experimental and control groups.

## Additional Formulas for the Standard Error of Effect Sizes

Hedges' Formula 9 would not have yielded an accurate standard error for any of the 14 effect sizes calculated by Kulik et al. Additional formulas are needed to calculate standard errors from studies comparing gain scores and from studies using analysis of covariance to measure treatment effects. Fortunately, it is easy to derive such formulas. To do so, one takes advantage of an algebraic relations between (a) the formulas for calculating average effect sizes for various experimental designs, and (b) the $t$ ratios used to test the statistical significance of the treatment effects found with these designs. The relationship between $t$ ratios and effect size formulas has already been demonstrated by Glass and his colleagues (Glass et al. 1981, chapter 5).

Effect size calculated from gains of independent groups. When an experimenter has estimated a treatment effect by comparing gains of independent experimental and control groups, the significance of the effect is usually tested by a $t$ ratio. When the assumptions for the use of this statistical test are met, the effect size in the experiment can be estimated by Glass's formula for comparing gains (Formula 2). Glass and his colleagues (Glass et al., 1981, p. 127) have shown that an effect size calculated in this way is equal to the following:

$$\underline{d}_G = \underline{t}_G \sqrt{2(1 - \rho_{\underline{XY}})(1/\underline{n}^E + 1/\underline{n}^C)} , \qquad (10)$$

where $\underline{t}_G$ is the $t$ ratio for testing a difference in gain scores and $\rho_{\underline{XY}}$ is the population correlation, ordinarily estimated by $r_{\underline{XY}}$, between the pretest ($\underline{X}$) and the posttest ($\underline{Y}$).

It is easy to see from Equation 10 that the effect size $\underline{d}_G$ is equal to a constant times the $t$ ratio. When assumptions for the use of this statistical test have been met, the sampling distribution of an unbiased estimator of this effect size is that of a noncentral $t$ times the constant. With both $\underline{n}^E$ and $\underline{n}^C$ large, the distribution of $\underline{d}_G$ will be approximately normal with $\mu(\underline{d}_G) = \delta$ and

$$\sigma(\underline{d}_{\underline{G}}) = \sqrt{2(1 - \rho_{\underline{XY}})\left[\frac{1}{\underline{n}^{E}} + \frac{1}{\underline{n}^{C}}\right] + \frac{\delta^2}{2(\underline{n}^{E} + \underline{n}^{C})}} \quad . \qquad (11)$$

When pretest and posttest correlate more than .5, the standard error of a study effect calculated from gains will be smaller than the standard error of the study effect calculated from the posttests only.

Effect sizes estimated from covariance-adjusted postscores of independent groups. Treatment effects are often tested for significance using analysis of covariance. When assumptions are met for use of the $\underline{t}$ ratio or $\underline{F}$ ratio for testing the difference between the covariance-adjusted means of the experimental and control groups, the effect size $\underline{d}_{\underline{g}}$ can be estimated by the following:

$$\underline{d}_{\underline{g}} = \frac{(\underline{M}_{\underline{Y}}^{E} - \underline{M}_{\underline{Y}}^{C}) - \underline{b}_{\underline{YX}}(\underline{M}_{\underline{X}}^{E} - \underline{M}_{\underline{X}}^{C})}{\underline{S}_{\underline{Y}}} \quad , \qquad (12)$$

where $\underline{b}_{\underline{YX}}$ is the pooled within-group estimate of the regression of final status $\underline{Y}$ on the covariate $\underline{X}$. Glass and his colleagues have shown that this effect size is equal to a constant times the $\underline{t}$ ratio calculated for the same design (Glass, 1981, p. 127).

$$\underline{d}_{\underline{g}} = \underline{t}_{\underline{g}} \sqrt{(1 - \rho_{\underline{YX}}^{2})\left[\frac{1}{\underline{n}^{E}} + \frac{1}{\underline{n}^{C}}\right]} \quad , \qquad (13)$$

where $\underline{t}_{\underline{g}}$ is the $\underline{t}$ ratio for testing the treatment effect for this experimental design and $\rho_{\underline{YX}}$ is the population correlation, ordinarily estimated by the same value $\underline{r}_{\underline{YX}}$, between the final-status score $\underline{Y}$ and the covariate $\underline{X}$. The sampling distribution of an unbiased estimator of this effect is that of a noncentral $\underline{t}$ times the constant. With both $\underline{n}^{E}$ and $\underline{n}^{C}$ large, the distribution of $\underline{d}_{\underline{g}}$ will be approximately normal with $\mu(\underline{d}_{\underline{g}}) = \delta$ and

$$\sigma(\underline{d}_{\underline{g}}) = \sqrt{(1 - \rho^2_{\underline{YX}})\left[\frac{1}{\underline{n}^E} + \frac{1}{\underline{n}^C}\right] + \frac{\delta^2}{2(\underline{n}^E + \underline{n}^C)}}, \qquad (14)$$

A comparison of Formulas 14 and 9 shows that standard errors of effect sizes calculated from residual scores are generally smaller than are standard errors of effect sizes calculated from posttests only.

Effect sizes estimated from posttest scores of matched groups. When a researcher has used a matched-pairs design to test the effect of an experimental treatment, the effect size can be calculated from Formula 1. With such a design, the statistical significance of the effect is usually tested with a $\underline{t}$ ratio for correlated means. Glass has shown that this $\underline{t}$ ratio is related to the effect size calculated for this design by the following (Glass et al., 1981, p. 127):

$$\underline{d}_{\underline{D}} = \underline{t}_{\underline{D}} \sqrt{\frac{2(1 - \rho_{\underline{YY}'})}{\underline{N}}}, \qquad (15)$$

where $\rho_{\underline{YY}'}$ is the population correlation of the paired posttest scores. This equation shows that an effect size estimated from this design, $\underline{d}_{\underline{D}}$, is equal to a constant times the $\underline{t}$ ratio calculated from the same design. The sampling distribution of an unbiased estimate of this effect size is that of a $\underline{t}$ ratio times the constant. With $\underline{N}$ large, the distribution of $\underline{d}_{\underline{D}}$ will be approximately normal $\mu(\underline{d}_{\underline{D}}) = \delta$ and

$$\sigma(\underline{d}_{\underline{D}}) = \sqrt{\frac{2(1 - \rho_{\underline{YY}'})}{\underline{N}} + \frac{\delta^2}{2\underline{N}}}. \qquad (16)$$

## Implications for Meta-analyses in the Literature

The lack of explicitness in recent books on meta-analysis about the need for a variety of formulas for effect sizes and standard errors has led to flawed analyses and flawed conclusions in the research literature. We focus here on three meta-analyses

that have used the newer statistical methods for research synthesis. The reviews were selected for special examination on the basis of the detail that they contain. The first and second reviews (Becker, 1983; Rosenthal & Rubin, 1982) list effect sizes and standard errors for individual studies. The third review (Pearlman, 1984) contains a numerical calculation of cumulated sampling error in a set of effect sizes listed in a report by Kulik et al. (1984). The three reviews contain enough detail for readers to reconstruct what was actually done in the analyses.

Interpersonal expectancies. To illustrate the utility of their method of research synthesis, Rosenthal and Rubin (1978, 1982) applied their test of homogeneity of effect sizes to findings on interpersonal expectancy reported originally in a dissertation by Keshock (1971). Subjects in Keshock's dissertation were 48 black inner city boys in grades 2 through 5. Half the children at each grade level were assigned to the experimental treatment, and their teachers were told that these children showed an ability level one standard deviation greater than their actual scores. The teachers of the control children were given the children's actual ability scores. Keshock reported that this experimental treatment had no significant effect on the children's achievement scores, as measured by the Wide-Range Achievement Test (WRAT).

Rosenthal and Rubin concluded that Keshock's analysis was inadequate because it failed to take into account grade-level differences in treatment effects. They reanalyzed Keshock's data, using the WRAT gain scores that Keshock listed in an appendix to his dissertation. Rosenthal and Rubin reported their conclusions succinctly:

> Gains in performance were substantially greater for the children whose teachers had been led to expect greater gains in performance. The sizes of the effects varied across the four grades from nearly half a standard deviation to nearly four standard deviations. For all subjects combined, the mean effect size was 2.04. (P. 383).

It is not at all obvious how Keshock could have missed an effect of this magnitude. Cohen (1977) has given rough guidelines for interpreting effect sizes. According to Cohen, effects of about 0.8 standard deviations should be considered large; they can usually be detected by eye without the aid of special measuring tools. Although Rosenthal and Rubin's reanalysis showed an average effect size of 2.04, Keshock had not noticed any effect of teacher expectations on WRAT scores. Nor did his original

statistical analysis disclose such effects. How could Keshock have failed to note so large an effect?

How also could Keshock's experimental treatment have produced an effect of this magnitude? Teachers of children in the experimental group were simply told that their children had IQs one standard deviation higher than their actual IQs. According to Rosenthal and Rubin, this simple manipulation raised achievement scores of experimental-group children by an average of two standard deviations. Scores of second graders rose by an average of almost four standard deviations. These are enormous gains: One standard deviation on an IQ scale is equal to 15 points, for example; a gain of two standard deviations on such a scale is equivalent to 30 points, and a gain of four standard deviations is equivalent to 60 points.

Rosenthal and Rubin's results are not so paradoxical as they may at first appear to be. Rosenthal and Rubin calculated effect sizes from Keshock's $t$ ratios for testing the significance of a difference in gains, but they did not use the appropriate formula (Equation 10). Instead, they converted the $t$ ratios to effect sizes using an equation appropriate for the $t$ test for comparing final-status scores:

$$d = t\sqrt{1/n^E + 1/n^C} \quad . \tag{17}$$

Rosenthal and Rubin therefore expressed the treatment effects for Keshock's study not in terms of variation in achievement but rather in terms of variation in achievement gains. Such effect sizes are not interpretable on the same scale as are other effect sizes.

Conventional effect sizes can be estimated easily for Keshock's experiment. Scores on the WRAT reading and arithmetic tests are standardized scores with a mean of 100 and a standard deviation of 15. Raw treatment effects reported by Keshock on the WRAT were 5.54 points in reading and 5.91 points in arithmetic. These gains are analogous to increases of this magnitude on an IQ scale. Just as it would be misleading to refer to a gain of 6 IQ points as representing 2.04 standard deviations, it is misleading to refer to a gain of this size on the WRAT as equivalent to an effect size of 2.04. The average effect size on the WRAT was 0.37 standard deviations in reading and 0.39 standard deviations in arithmetic.

Table 1 compares our estimates of effect sizes for Keshock's data with estimates made by Rosenthal and Rubin. Our calculations

of composite effect sizes are based on a reported correlation of .7 between arithmetic and reading scores on the WRAT. Our estimated standard errors are based on retest correlations of .93 for WRAT composite scores.

Gender and susceptibility to influence. Becker's (1983) effect size synthesis on this topic used Hedges' (1982) homogeneity approach. More than 100 studies in this area were originally located by Eagly, who used both a box-score approach and meta-analytic methodology to integrate the results of the studies (Eagly, 1978; Eagly & Carli, 1981). Eagly's analyses showed that males and females differed significantly in susceptibility to influence in the three areas of persuasion, conformity, and group pressure. Becker thought, however, that a reanalysis of Eagly's data was needed. To her, Eagly's box-score approach seemed clearly inadequate, and even Eagly and Carli's meta-analytic methods seemed to be ad hoc.

Becker examined individually each of the effect sizes reported by Eagly and Carli, and she reported that they were quite accurate for the most part. Only a handful of the nearly 100 effect sizes recalculated by Becker differed by as much as 0.10 from Eagly and Carli's original estimates. Becker used her own recoded effect sizes rather than Eagly and Carli's estimates in her reanalysis of results in this area.

Becker's reanalysis led her to question Eagly and Carli's conclusions. Like Eagly and Carli, Becker found average differences between males and females in susceptibility to influence, but she did not attach much weight to this finding. More important to her was her observation that variation in study effects was more closely related to study methodology than it was to indicators of sex bias. Becker pointed out, for example, that variation in results in 36 persuasion studies was related to type of outcome variable used in the study--a methodological feature--rather than to such factors as gender of the investigator, gender bias in message, and so on. Studies that used postscores on an attitude measure as the outcome variable (Group I studies) produced near-zero effects; studies that used change scores, covariance-adjusted scores, or their equivalent (Group II studies) produced more sizeable effects.

It is possible to evaluate Becker's conclusions because she provided a figure showing effect sizes and standard errors of these effects for the 36 persuasion studies. Comparing Becker's statistics with results in the original reports shows that she calculated operative effect sizes for all studies--no matter what type of experimental design or dependent variable was used in a

study. Such effect sizes do not estimate the same quantities for different research designs.

Becker's calculations led her to reach the following conclusion:

> Methodological considerations rather than features representing sex bias explain the variability in persuasion study results. The more stable sex difference for persuasion studies that had been noted by Maccoby and Jacklin (1974) seems to be largely spurious. Given that we can account for the size of the differences with this methodological artifact, claims of sex bias must hold less sway. (P. 13).

It seems to us that it is not a methodological characteristic of the studies that explains the different effect sizes in Group I and Group II experiments. It is rather Becker's method of calculating effect sizes that explains her finding. Group I treatment effects were standardized on a final-status measure, $S_Y$; Group II treatment effects were standardized on measures of gain, $S_G = S_Y\sqrt{2(1 - r_{XY})}$, or on residual measures, $S_g = S_Y\sqrt{1 - r_{XY}^2}$. Given the difference in the units on which Group I and II treatment effects are standardized, meaningful conclusions cannot be drawn from an analysis of the combined data set.

Coaching effects. Pearlman's (1984) effect size synthesis was based on 38 studies of coaching effects originally analyzed by Kulik et al. (1984). Kulik and his colleagues had concluded from their meta-analysis of results reported in the studies that coaching programs in general have positive effects on aptitude test performance. They cautioned, however, that results differed in two literatures on coaching: the literature on the SAT and the literature on other aptitude tests. Kulik and his co-workers were unable to explain much of the variability in effect sizes within these two literatures, and they concluded that "it was impossible to explain fully why coaching results differ from study to study as much as they do" (p. 187).

Pearlman analyzed the data assembled by Kulik and his co-workers using the methodology for effect size synthesis developed by Hunter et al. (1982). Pearlman's analysis led to conclusions that differed from Kulik's on all major points. Pearlman concluded, for example, that coaching effects are small for both the SAT and for other tests. Most important, Pearlman concluded that a substantial portion of the observed variance in effect sizes was attributable to a statistical artifact--sampling error--

rather than to true differences in results from different studies. Pearlman interpreted his results as supporting the conclusion that between-study differences in coaching effectiveness are much less extensive than they appear to be.

Close examination of Pearlman's calculations show that they are seriously flawed. These flaws can be seen most clearly in Pearlman's analysis of results from 14 SAT studies included in the total pool of 38 studies. Pearlman calculated the total sampling error in the 14 studies using a cumulation formula (Hunter et al., 1982, p. 102) that is closely related to Formula 9 and is appropriate for calculating sampling error with independent-group, posttest-only experimental designs. None of the 14 SAT studies used such a design, however, and none of the 14 effect sizes for these studies was calculated from such formulas as 1 or 17. The formulas that Kulik and his co-workers used to estimate effect sizes were formulas for interpretable effect sizes: Formulas 2, 10, 12, 13, and 15. The appropriate formulas for calculating standard errors of these effect sizes are Formulas 11, 14, and 16.

The results of Pearlman's failure to take experimental design into account are serious. He estimated that 51 per cent of the variance in the distribution of SAT study effects could be attributed to sampling errors in the individual studies. If he had taken into account the fact that all 14 SAT studies used pre-post designs and that SAT retests correlate .88, he would have had to conclude that sampling error of effect sizes in individual studies could not account for more than 12 per cent of the variation in study effects. Because some of the SAT studies used additional covariates, matched-pairs designs, and factorial designs that further reduced sampling errors, the true proportion of study effect variation attributable to sampling error may be even lower.

## Conclusions

Although many researchers have found Glass's meta-analytic methodology appealing and useful, some methodologists have criticized the statistics that underlie this methodology. Hedges and Olkin (1982), for example, have described Glass's procedures as ad hoc and generally inappropriate. They have also asserted that hundreds of meta-analyses patterned on Glass's model have used statistics that are of questionable validity or that are demonstrably incorrect (Hedges & Olkin, 1985, p. 14). "The conclusions of these meta-analyses may indeed be correct," Hedges and Olkin have written, "but the statistical reasoning in support of these conclusions is not" (1985, p. 14).

Hedges and Olkin (1985), Hunter, Schmidt, and Jackson (1982), and Rosenthal (1984) have in recent books tried to firm up the statistical basis of meta-analysis. Careful examination of their books shows, however, that they fail to distinguish between interpretable and operative effect sizes, and they ignore Glass's guidelines on the calculation of interpretable effect sizes for studies with different experimental designs. Their failure to consider the influence of experimental design on effect size calculation seriously limits the utility of their work.

It is no surprise to find that users of the statistics advocated in these recent books have produced works with serious flaws:

1. These meta-analyses have often been based on operative rather than interpretable effect sizes. Because operative effect sizes are usually calculated with reduced standardization terms, operative effect sizes are inappropriate for use in meta-analysis. When operative effect sizes rather than interpretable ones are used in research syntheses, an artifactual relationship emerges between effect sizes and experimental designs, and average effect sizes often become seriously inflated.

2. These meta-analyses have often been based on miscalculated standard errors. Methodologists who have written about standard errors of effect sizes have presented only one formula for calculating such standard errors. This formula gives reasonable results only when applied to studies using an unblocked, posttest-only design. The formula produces inaccurate results when applied to studies that estimate effects with greater precision--or for the majority of studies in the social sciences.

We believe that valid conclusions cannot be drawn from meta-analyses in which effect sizes are miscalculated. Nor can valid conclusions be drawn when meta-analysts use inflated standard errors to test the homogeneity of collections of effect sizes. We are therefore more pessimistic in our assessment of recent meta-analytic work than Hedges and Olkin were in their evaluation of earlier meta-analytic results. We believe that both the statistical reasoning and the conclusions are likely to be incorrect in studies that have used the newer statistical methods for research synthesis.

## References

Bangert-Drowns, R. L. (in press). A review of developments in meta-analytic method. Psychological Bulletin.

Becker, B. J. (1983, April). Influence again: A comparison of methods for meta-analysis. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

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

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. (Rev. ed.). New York: Academic Press.

Eagly, A. H. (1978). Sex differences in influenceability. Psychological Bulletin, 85, 86-116.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sextyped communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. Psychological Bulletin, 90, 1-20.

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

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

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

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92(2), 490-499.

Hedges, L. V., & Olkin, I. (1982). Analyses, reanalyses, and meta-analysis. Contemporary Education Review, 1, 157-165.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Advanced meta-analysis: Quantitative methods for cumulating research findings across studies. San Francisco: Sage Publications.

Keshock, J. D. (1971). An investigation of the effects of the expectancy phenomenon upon the intelligence, achievement and motivations of inner-city elementary school children. Dissertation Abstracts International, 32, 01-A (University Microfilms No. 71-19,010)

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

Kulik, J. A., Bangert, R. L., & Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. Psychological Bulletin, 95, 179-188.

McGaw, B., & Glass, G. V. (1980). Choice of the metric for effect size in meta-analysis. American Educational Research Journal, 17, 325-337.

Pearlman, K. (1984, August). Validity generalizations: Methodological and substantive implications for meta-analytic research. Paper presented at the annual meeting of the American Psychology Association, Toronto, Canada.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills, CA: Sage Publications.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. The Behavioral and Brain Sciences, 1, 377-386.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92(2), 500-504.

## Author Note

The material in this report is based upon work supported by National Science Foundation Grant No. MDR 8470258. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Requests for reprints should be sent to James A. Kulik, Center for Research on Learning and Teaching, The University of Michigan, 109 E. Madison St., Ann Arbor, Michigan 48109.

Table 1

Comparison of Effect Sizes in Keshock's (1970) Study Estimated by Different Formulas

| Grade | Effect Size Estimated by Rosenthal & Rubin (1982) | | Effect Size Estimated from Formulas 2 and 11 | |
|---|---|---|---|---|
| | M | SD | M | SD |
| 2 | 3.85 | 1.03 | 0.89 | 0.22 |
| 3 | 2.34 | 0.78 | 0.53 | 0.22 |
| 4 | 0.47 | 0.58 | 0.11 | 0.22 |
| 5 | 1.48 | 0.66 | 0.34 | 0.22 |
| Mean | 2.04 | 0.73 | 0.44 | 0.11 |