Psychological Bulletin
1982, Vol. 92, No. 2, 490-499

0033-2909/82/9202-0490$00.75

Estimation of Effect Size From a Series of Independent Experiments

Larry V. Hedges
Department of Education, University of Chicago

Recent interest in methods for the quantitative synthesis of research has produced many quantitative syntheses but little statistical theory for the methods used in these syntheses. The present article extends statistical theory for procedures based on Glass's estimator of effect size. An unbiased estimator of effect size is given. A weighted estimator of effect size based on data from several experiments is defined and shown to be optimal (asymptotically efficient). An approximate (large-sample) test for homogeneity of effect size across experiments is also given. The results of an empirical sampling study are used to show that the large-sample distributions of the weighted estimator and the homogeneity statistic are quite accurate when the experimental and control group sample sizes exceed 10 and the effect sizes are smaller than about 1.5.

A number of authors have recently shown an interest in empirical methods of combining the results of a series of independent studies. Glass (1976) was among the first authors to call for the use of quantitative procedures in research integration as a supplement to the discursive review. Examples of the use of these techniques include a review of psychotherapy outcome studies (Smith & Glass, 1977), a review of interpersonal expectancy effects in psychological research (Rosenthal & Rubin, 1978), a review of the effects of teaching methods (Kulik, Kulik, & Cohen, 1979), and a review of the effects of class size on academic achievement (Glass & Smith, 1979).

Despite the extensive application of the techniques suggested by Glass, there has been relatively little attention to some fundamental statistical issues in the use of quantitative methods for research synthesis. Investigators have not always clearly distinguished sample *estimates* of effect size from the population parameter they wish to estimate. There has also been little discussion of whether it is reasonable to think that population effect sizes are constant across a series of studies. Given a series of effect size estimates from k studies, most investigators calculate the average of these effect size estimates and refer to the average value as *the* effect size. 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. 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.

When a series of studies seems to have a common population effect size, there is the question of the best procedure for estimating that common effect size. All investigators seem to start by obtaining an effect size estimate for each study using Glass's estimator of effect size. Some investigators have advocated the use of a combined estimator weighted by sample size. Other investigators combine estimates from individual studies without weighting. Neither choice seems to have been supported by formal statistical reasoning.

There has not been a great deal of literature on statistical aspects of research syn-

This research was supported by the Spencer Foundation.

I thank Betsy Jane Becker for her helpful comments and for programming the simulation study reported in the article.

Requests for reprints should be sent to Larry V. Hedges, Department of Education, University of Chicago, 5835 S. Kimbark Avenue, Chicago, Illinois 60637.

thesis using estimates of effect size. Glass (1976, 1978) proposed estimation of effect size but presented no sampling theory for his procedures. Hedges (1981) obtained the distribution of Glass's estimator of effect size and used that distribution to show that Glass's estimator is biased. Hedges also obtained an unbiased estimator of effect size that has smaller variance than Glass's estimator.

The present article extends Hedges' results by considering the statistical problem of estimating effect size from a series of experiments. A model for the results of a series of experiments is given, and some estimators of effect size for a single experiment are discussed. A large-sample test for the equality of k effect sizes is also given. This approximate test can be used as an indication of whether it is reasonable to pool the effect size estimates from a series of studies. A procedure for pooling estimates is given, and this procedure is shown to yield a pooled estimator with the same asymptotic distribution as the maximum likelihood estimator of effect size. Hence, the pooled estimator is asymptotically efficient. The validity of the large sample results is investigated with an empirical sampling study. Finally, the methods presented in this article are applied to some data from educational psychology.

Assumptions and Model

Earlier treatments of effect size have not adequately emphasized the assumptions underlying effect size estimation and testing. Cohen (1977) proposed the measure d of effect size in connection with the t test for the difference between means. Glass (1976) proposed the quantitative synthesis of the results of a collection of experimental/control group studies by estimating d 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 d 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 a control group (C). Let Y_{ij}^{E} and Y_{ij}^{C} be the jth scores on the ith 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 μ_i^{E} and μ_i^{C} and common variance σ_i^2, that is,

$$Y_{ij}^{E} \sim \mathcal{N}(\mu_i^{E}, \sigma_i^2),$$

$$j = 1, \ldots, n_i^{E}, \quad i = 1, \ldots, k,$$

and

$$Y_{ij}^{C} \sim \mathcal{N}(\mu_i^{C}, \sigma_i^2),$$

$$j = 1, \ldots, n_i^{C}, \quad i = 1, \ldots, k.$$

In this notation, the *effect size* for the ith study (δ_i) is defined as

$$\delta_i = \frac{\mu_i^{E} - \mu_i^{C}}{\sigma_i}, \tag{1}$$

where we use the Greek letter δ (instead of d) to denote that this effect size is a *population* parameter.

Note that the effect size δ is invariant under linear transformation of the observations. The values of the population means and the standard deviations change under linear transformations, but the effect size remains the same. This implies that if the same population of test scores is represented on two tests that are linearly equatable, the effect sizes will be identical. The virtue of effect sizes is that they are comparable even though they may be derived from different but linearly equatable measures.

Conversely, if two measures are not linearly equatable, then the same population of test scores would in general yield different effect sizes when represented on the two measures. In particular, if two tests do not measure the same construct, there is little reason to believe that effect sizes based on those two tests would be the same. The implication is that even if a treatment produces a uniform effect size on measures on one construct (such as mathematics achievement), there is little reason to expect that effect size to be the same as the effect size for studies that measure the influence of the same treatment on another construct (such as attitude).

Estimating Effect Size

The definition of effect size given in (1) above defines a population parameter δ_i in terms of other population parameters μ_i^E, μ_i^C, and σ_i. We will seldom, if ever, know the exact values of μ_i^E, μ_i^C, and σ_i; thus, we will have to *estimate* δ_i. Glass (1976) proposed a statistic g_i' to estimate δ_i by essentially replacing μ_i^E, μ_i^C, and σ_i in the definition of δ_i by their sample analogues. Specifically, Glass proposed the estimator g_i' of δ_i, where g_i' is defined by

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

where $\bar{Y}_i^E$ and $\bar{Y}_i^C$ are the experimental and control group sample means for the ith study and S_i^C is the control group sample standard deviation. Hedges (1981) showed that under the assumptions of the previous section, the estimator (2) is biased and that a less biased estimator results when S_i^C is replaced with the usual pooled within-groups standard deviation. We denote this estimator by g_i, that is,

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

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 g_i is a *sample statistic* and therefore has a sampling distribution of its own. Our assumptions imply that g_i is distributed as $(1/\sqrt{\tilde{n}_i})$ times a noncentral t random variable with $n_i^E + n_i^C - 2$ *df* 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 δ_i.

An Unbiased Estimator of Effect Size

A simple unbiased estimator of δ was obtained by Hedges (1981) based on the assumptions of the previous section. The unbiased estimator g_i^U is given by

$$g_i^U = c(m)g_i, \qquad (4)$$

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 defined, for example, in Mood, Graybill, and Boes (1974), and $c(m)$ is given approximately by

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

It is clear that as m becomes large, g_i^U tends to g_i, so g_i is almost unbiased in large samples. Because $c(m) < 1$, the variance of the unbiased estimator g_i^U is always smaller than the variance of g_i. Hence, g_i^U has uniformly smaller mean squared error than does g_i. The exact variance of g_i^U is

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

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

Asymptotic Distribution of the Unbiased Estimator

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

$$g_i^U \sim \mathcal{N}[\delta_i, \sigma_i^2(\delta_i)], \qquad (7)$$

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^U depends on the true effect size δ_i. This large-sample approximation is used by substituting an estimator of the effect size for δ_i in (8). In the case of a single effect size, we substitute g_i^U for δ_i in (8) to obtain an expression for the variance of g_i^U.

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

$$H_0: \delta_i = \delta, \quad i = 1, \ldots, k$$

versus the alternative that at least one δ_i differs from the rest.

A large-sample (approximate) test for the equality of k effect sizes uses the test statistic

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

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

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

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

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, the asymptotic distribution of H can be used as the basis for an approximate test of the homogeneity of the δ_i. (See the Appendix.) If all the k studies have the same population effect size (i.e., if H_0 is true), then the test statistic H has an asymptotic chi-square distribution given by

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

Therefore, if the obtained value of H exceeds the $100(1 - \alpha)$ % critical value of the chi-square distribution with $(k - 1)$ *df*, we reject the hypothesis that the δ_i are equal. If we reject this null hypothesis we may decide not to pool all of the estimates of δ because 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^U because rather small differences may lead to large values of the test statistic. If the g_i^U 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.

Estimation of Effect Size From a Series of Homogeneous Studies

If a series of k independent studies share a common effect size δ, it is natural to estimate δ 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

$$\sum_{i=1}^{k} w_i g_i^U, \tag{11}$$

where $w_i > 0$, $i = 1, \ldots, k$, 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

$$w_i = \frac{1/v_i}{\sum_{j=1}^{k} 1/v_j}, \quad i = 1, \ldots, k, \tag{12}$$

where v_i is the variance of g_i^U given in (6). The practical problem in calculating the most precise weighted estimate is that the ith weight depends on the variance of g_i^U, which in turn depends on δ.

One approach to the problem of weighting is to use weights that are based on some approximation to the v_i that does not depend on δ. 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 δ is near zero these weights will be close to optimal because v_i depends on δ^2, which will be small. If a nonzero a priori estimate of δ is available, then weights could be estimated by inserting that value of δ in expression (6) for the variance of g_i^U and using the formula (12) for w_i. In general, the result will be an unbiased pooled estimator of δ that is slightly less precise than the most precise weighted estimator.

Estimating Weights

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

$$g. = \frac{\sum_{i=1}^{k} \frac{g_i^U}{\sigma_i^2(g_i^U)}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2(g_i^U)}}, \qquad (13)$$

where $\sigma_i^2(\delta_i)$ is given by (8). The estimator $g.$ is therefore obtained by calculating the weights using g_i^U for δ_i in (8). Although the g_i^U 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^U by $g.$ in the expression for $\sigma_i^2(g_i^U)$ and iterating. That is, calculate the estimator $g.^{(1)}$ defined by

$$g.^{(1)} = \frac{\sum_{i=1}^{k} \frac{g_i^U}{\sigma_i^2(g.)}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2(g.)}}, \qquad (14)$$

where $\sigma_i^2(\delta_i)$ is given by (8). The iterated estimator $g.^{(1)}$ 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 δ based on $g.$. The 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. (See the Appendix.) The large sample approximation is

$$g. \sim \mathcal{N}[\delta, \sigma.^2(\delta)], \qquad (15)$$

where

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

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

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

If the obtained value of $z(g.)$ is larger in absolute value than the $100(1 - \alpha/2)$ % critical value of the standard normal distribution, we reject the hypothesis that $\delta = 0$ at the 100α % 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 δ in (16). Therefore, confidence intervals and significance tests for δ 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. The answer to this question is that $g.$ is asymptotically efficient in the sense that the asymptotic variance of $g.$ is the theoretical minimum (Cramér-Rao bound). Thus, no other con-

sistent estimator has smaller asymptotic variance. This result implies that $g.$ has the same asymptotic distribution as the maximum-likelihood estimator of δ based on k experiments.

Accuracy of the Large-Sample Approximation

The statistical procedures described in this article depend on large-sample approximations to the distributions of g_i^U, $g.$, and $H.$ Although large-sample approximations are sometimes reasonably accurate in small samples, the uncritical use of large-sample statistical theory is unjustified. The asymptotic theory used in this article is correct for any fixed δ, but we would expect this asymptotic theory to be most accurate in small samples when δ is small. In order to evaluate the accuracy of the large-sample approximations used here, a simulation study was conducted. All the simulations described in this section are based on standard normal deviates and chi-squared random numbers generated by the International Mathematical and Statistical Libraries (1977) subroutines GGNML and GGCHS. In each simulation four representative effect sizes were used: $\delta = .25$, $\delta = .50$, $\delta = 1.00$, and $\delta = 1.50$.

Table 1
Small-Sample Accuracy of Confidence Intervals for δ Based on the Normal Approximation to the Distribution of g^U

Sample size $n = n^E = n^C$	Mean of g^U	Variance of g^U	Proportion of confidence intervals containing δ with nominal significance level					
			.60	.70	.80	.90	.95	.99
			$\delta = .25$					
10	.252	.20631	.621	.714	.813	.910	.955	.991
20	.255	.10291	.604	.704	.806	.903	.951	.990
30	.246	.06730	.608	.708	.809	.908	.954	.989
40	.248	.05138	.601	.703	.800	.900	.950	.991
50	.250	.03946	.612	.709	.809	.904	.952	.991
100	.251	.02028	.600	.705	.808	.903	.949	.990
			$\delta = .50$					
10	.504	.21386	.620	.714	.807	.906	.954	.990
20	.497	.10452	.609	.709	.807	.904	.955	.990
30	.500	.06977	.603	.699	.800	.903	.952	.990
40	.499	.05225	.599	.700	.803	.903	.952	.991
50	.496	.04194	.597	.696	.799	.904	.953	.990
100	.498	.02052	.606	.698	.799	.906	.954	.989
			$\delta = 1.00$					
10	.993	.24119	.602	.703	.801	.901	.952	.992
20	.993	.11202	.609	.707	.808	.907	.955	.992
30	.993	.07754	.594	.697	.799	.897	.952	.991
40	.995	.05683	.607	.709	.805	.901	.953	.992
50	.996	.04604	.603	.706	.806	.900	.951	.989
100	1.000	.02252	.607	.702	.805	.905	.953	.990
			$\delta = 1.50$					
10	1.498	.28182	.599	.696	.797	.901	.953	.990
20	1.501	.13707	.601	.697	.797	.898	.950	.989
30	1.505	.09090	.600	.699	.797	.899	.950	.991
40	1.502	.06884	.594	.693	.796	.899	.949	.990
50	1.500	.05572	.593	.691	.794	.897	.949	.991
100	1.500	.02606	.603	.705	.810	.906	.955	.990

Note. Data for $n = 100$ are based on 4,000 replications. All other figures are based on 10,000 replications.

These effect sizes were chosen because virtually all meta-analyses have found effect sizes in this range. One should be cautious, however, about extrapolating the results of this (or any) simulation beyond the actual range of parameters studied. In each simulation, the experimental and control group sample sizes were set equal, that is, $n_i^E = n_i^C$. The g_i values were generated based on the identity

$$g = X/\sqrt{S/m},$$

where X is a normal with mean δ and variance $2/n$ and S is a chi-square random variable with $m = 2n - 2$ *df*. The g values were then transformed into g^U values using (4).

The large-sample normal approximation (7) underlies the other large-sample approximations used in this article. The accuracy of this approximation was studied by generating g_i^U values and using the approximation (7) to construct confidence intervals. The empirical accuracy of those confidence intervals is reported in Table 1. It is clear that the large-sample approximation (7) is quite accurate for sample sizes as small as $n^E = n^C = 10$. It also appears that the accuracy of the approximation is satisfactory throughout the range of effect sizes studied.

The distributions of the weighted estimators $g.$ given in (13) and $g.^{(1)}$ given in (14) were also investigated. The cases $k =$

Table 2
Small-Sample Accuracy of Confidence Intervals for δ Based on the Normal Approximation to the Distribution of g. and g.$^{(1)}$

	Summary of empirical proportions of confidence intervals containing δ for nominal significance level α								
	$1 - \alpha = .90$[a]			$1 - \alpha = .95$[b]			$1 - \alpha = .99$[c]		
δ	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
				Estimator g. $k = 2$					
.25	.898	.908	.920	.944	.953	.964	.989	.990	.994
.50	.891	.903	.919	.943	.954	.962	.988	.990	.995
1.00	.891	.904	.914	.946	.954	.960	.988	.991	.995
1.50	.890	.903	.917	.945	.950	.959	.988	.991	.992
				Estimator g. $k = 5$					
.25	.894	.906	.922	.946	.954	.964	.987	.991	.993
.50	.894	.903	.915	.938	.950	.961	.986	.9895	.993
1.00	.883	.895	.917	.937	.947	.957	.987	.989	.994
1.50	.886	.900	.906	.941	.948	.952	.984	.991	.993
				Estimator g.$^{(1)}$ $k = 2$					
.25	.893	.904	.912	.942	.951	.957	.988	.990	.992
.50	.888	.902	.912	.943	.952	.960	.987	.990	.995
1.00	.891	.901	.913	.939	.951	.957	.987	.990	.995
1.50	.884	.898	.904	.942	.947	.954	.986	.990	.993
				Estimator g.$^{(1)}$ $k = 5$					
.25	.894	.902	.911	.940	.950	.959	.984	.990	.993
.50	.889	.899	.908	.936	.946	.958	.985	.988	.993
1.00	.881	.896	.911	.938	.946	.957	.985	.989	.993
1.50	.878	.896	.903	.938	.948	.959	.980	.990	.992

Note. Sample sizes for each estimate were $n_i^E = n_i^C = 10$, 20, 30, 40, or 50. For $k = 2$, 15 combinations of sample sizes were used. For $k = 5$, 20 configurations of sample sizes were used. For each configuration of sample sizes, 2,000 replications were generated.

[a] The standard error of each estimated proportion is approximately .0067.

[b] The standard error of each estimated proportion is approximately .0049.

[c] The standard error of each estimated proportion is approximately .0022.

Table 3
Small-Sample Behavior of the Homogeneity Test Statistic H

	Summary of empirical proportions of text statistics exceeding the nominal significance level α								
	$\alpha = .10$[a]			$\alpha = .05$[b]			$\alpha = .01$[c]		
δ	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
					$k = 2$				
.25	.082	.096	.110	.040	.048	.056	.005	.009	.012
.50	.083	.092	.114	.041	.047	.058	.007	.011	.014
1.00	.090	.098	.106	.041	.048	.056	.002	.009	.013
1.50	.093	.099	.109	.043	.049	.056	.006	.010	.016
					$k = 5$				
.25	.079	.094	.107	.035	.044	.056	.004	.008	.012
.50	.088	.096	.108	.040	.048	.057	.005	.009	.012
1.00	.083	.090	.098	.037	.046	.051	.005	.008	.014
1.50	.089	.105	.112	.041	.052	.057	.007	.010	.017

Note. Sample sizes for each estimate were $n_i^E = n_i^C = 10$, 20, 30, 40, or 50. For $k = 2$, 15 combinations of sample sizes were used. For $k = 5$, 20 configurations of sample sizes were used. For each configuration of sample sizes, 2,000 replications were generated.

[a] The standard error of each estimated proportion is approximately .0067.

[b] The standard error of each estimated proportion is approximately .0049.

[c] The standard error of each estimated proportion is approximately .0022.

2 and $k = 5$ (two or five independent studies) were studied extensively.

For $k = 2$, all 15 possible combinations of the five sample sizes ($n_i^E = n_i^C = 10$, 20, 30, 40, and 50) were studied. For $k = 5$, 20 different configurations of the sample sizes were studied. The most extreme configurations were $n_i^E = n_i^C = 10$ for all five estimators and $n_i^E = n_i^C = 50$ for all five estimators. A total of 2,000 replications of each sample size configuration were generated for each population effect size. The large-sample approximation was used to calculate confidence intervals for δ. Means and variances of the estimates as well as the empirical proportion of confidence intervals that contained δ were calculated for each sample size configuration for each effect size.[1] The results of these simulations suggest that the large-sample approximation to the distribution of the estimators is reasonably accurate for the range of δ examined, even when all the studies have a sample size of 10 per group. The accuracy of the approximation tends to improve as the sample sizes increase. As expected, the estimator $g_.$ has a slight negative bias, tending to underestimate δ. The simulation verifies that the iterated estimator $g_.^{(1)}$ is less biased than $g_.$, although neither estimator is markedly superior to the other. A summary of the empirical proportions of confidence intervals that contained δ is given in Table 2.

The accuracy of the chi-square test for homogeneity of effect size was also studied. The distribution of the statistic H was studied for $k = 2$ and $k = 5$ for the same combinations of sample sizes as were used in the study of $g_.$ and $g_.^{(1)}$. The large sample approximation to the distribution was also fairly accurate even when $n^E = n^C = 10$. The actual significance values of H have a slight tendency to be lower than the nominal significance levels. Thus, the test for homogeneity may be slightly conservative. The large sample distribution of H is also more accurate as δ and the sample sizes increase. Table 3 is a summary of the results on the empirical distribution of H.

[1] A complete report of the results of the simulation study is available from the author.

Table 4
Data for the Example: The Effects of Open (O) Versus Traditional (T) Teaching on Student Curiosity

Study	n						
	O	T	g_i^U	$\sigma_i^2(g_i^U)$	$\sigma^2(g_.)$	$\hat{w}_i^a$	$\hat{w}_i^b$
1	16	11	.459	.1573	.1537	.081	.082
2	30	30	.181	.0672	.0668	.190	.188
3	30	30	−.521	.0689	.0668	.185	.188
4	44	40	.097	.0478	.0478	.267	.263
5	37	55	.425	.0462	.0453	.276	.279

[a] Weight based on estimating δ by g_i^U.
[b] Weight based on estimating δ by $g_.$.

Example

The techniques described here were applied to some studies in research on teaching by Hedges, Giaconia, and Gage (Note 1). Some data from five studies that examined the effects of open versus traditional teaching on student curiosity are given in Table 4. Sample sizes, the effect size estimate, and weights are given for each study. The homogeneity statistic H has the value $H = 9.02$. Comparing this value with the 95% critical value (9.49) of the chi-square distribution with 4 df, we see that we cannot reject homogeneity of effect sizes for these five studies.

The weighted estimates of δ are $g_. = .119$ and $g_.^{(1)} = .118$. A 95% confidence interval for δ based on $g_.$ is

$$-.103 \leq \delta \leq .339$$

and a 95% confidence interval based on $g_.^{(1)}$ is

$$-.102 \leq \delta \leq .338.$$

Because these confidence intervals contain zero, we cannot reject the hypothesis that $\delta = 0$ at the $\alpha = .05$ level. This example illustrates that $g_.$ and $g_.^{(1)}$ are usually quite similar as are confidence intervals derived from the two estimators.

Reference Note

1. Hedges, L. V., Giaconia, R. M., & Gage, N. L. *The empirical evidence on the effects of open education* (Final report of the Stanford Research Synthesis Project, Vol. 2). Stanford, Calif.: Stanford University, School of Education, 1981.

References

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

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* (Vol. 5). Itasca, Ill.: Peacock, 1978.

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

International Mathematical and Statistical Libraries, Inc. *IMSL Library 1* (7th ed.). Houston, Tex.: Author, 1977.

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

Mood, A. M., Graybill, F. A., & Boes, D. C. *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill, 1974.

Rao, C. R. *Linear statistical inference and its applications*. New York: Wiley, 1973.

Rosenthal, R., & Rubin, D. B. Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1978, *3*, 377–415.

Rosenthal, 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.

Appendix

Formal Statement of Results Used in This Paper

The distribution of H and $g.$ can be obtained as direct consequences of a theorem on the large-sample distribution of weighted estimators (see, e.g., Rao, 1973, pp. 389–390). A formal statement of the theorem requires a regularity condition that the n_i^E and n_i^C, $i = 1, \ldots, k$ tend to infinity at the same rate. The actual results used in this article follow.

Result 1

If H_0 is true, $N = \sum_{i=1}^{k} (n_i^E + n_i^C)$, $\pi_i^E = n_i^E/N$, $\pi_i^C = n_i^C/N$, $i = 1, \ldots, k$, and π_i^E, π_i^C remain fixed as $N \to \infty$, then the asymptotic distribution of H given in (9) is χ^2_{k-1}.

Result 2

If $N = \sum_{i=1}^{k} (n_i^E + n_i^C)$, $\pi_i^E = n_i^E/N$, $\pi_i^C = n_i^C/N$, $i = 1, \ldots, k$, and π_i^E, π_i^C remain fixed as $N \to \infty$, then the asymptotic distribution of $g.$ defined in (13) is given by

$$\sqrt{N}(g. - \bar{\delta}.) \sim \mathcal{N}(0, \tilde{\sigma}.^2),$$

where

$$\tilde{\sigma}.^{-2} = \sum_{i=1}^{k} \frac{1}{\tilde{\sigma}_i^2(\delta_i)},$$

$$\tilde{\sigma}_i^2(\delta_i) = \frac{\pi_i^E + \pi_i^C}{\pi_i^E \pi_i^C} + \frac{\delta_i^2}{2(\pi_i^E + \pi_i^C)},$$

$$\bar{\delta}. = \frac{\sum_{i=1}^{k} \frac{\delta_i}{\tilde{\sigma}_i^2(\delta_i)}}{\sum_{i=1}^{k} \frac{1}{\tilde{\sigma}_i^2(\delta_i)}}.$$

A straightforward calculation shows that $\tilde{\sigma}.^2$ is the minimum asymptotic variance. Write the observations of the ith study in terms of δ, a location (scale mean) parameter γ_i, and a residual ϵ, that is,

$$Y_{ij}^E = \delta\sigma_i + \gamma_i + \epsilon_{ij}^E,$$

$$j = 1, \ldots, n_i^E, \quad i = 1, \ldots, k,$$

$$Y_{ij}^C = \gamma_i + \epsilon_{ij}^C, \quad j = 1, \ldots, n_i^C, \quad i = 1, \ldots, k,$$

where ϵ_{ij}^E and ϵ_{ij}^C are distributed independently as $\mathcal{N}(0, \sigma_i^2)$. The second derivatives of the likelihood yield a $(2k + 1) \times (2k + 1)$ matrix. The expectation of this matrix can be inverted using standard formulas for inversion of a patterned matrix. After simplification, the entry corresponding to the minimum asymptotic variance of a consistent estimator of δ is $\tilde{\sigma}.^2$. Because this is also the asymptotic variance of the maximum-likelihood estimator, we deduce that $g.$ has the same asymptotic distribution as the maximum-likelihood estimator.

Received October 5, 1981 ■