Journal of Educational Statistics
Spring 1984, Volume 9, Number 1, pp. 61–85

ESTIMATION OF EFFECT SIZE UNDER NONRANDOM SAMPLING: THE EFFECTS OF CENSORING STUDIES YIELDING STATISTICALLY INSIGNIFICANT MEAN DIFFERENCES

LARRY V. HEDGES
The University of Chicago

KEY WORDS. *Meta-analysis, effect size, research synthesis, censoring, publication bias, nonrandom sampling*

ABSTRACT. Quantitative research synthesis usually involves the combination of estimates of the standardized mean difference (effect size) derived from independent research studies. In some cases, effect size estimates are available only if the difference between experimental and control group means is statistically significant. If the quantitative result of a study is observed only when the mean difference is statistically significant, the observed mean difference, variance, and effect size are biased estimators of the corresponding population parameters. The exact distribution of the sample effect size is derived for the case in which only studies yielding statistically significant results may be observed. The maximum likelihood estimator of effect size also is derived under the model in which only significant results are observed. The exact distribution of the maximum likelihood estimator is obtained numerically and is used to study the bias of the maximum likelihood estimator. An empirical sampling study is used to supplement the analytic results.

Quantitative research synthesis usually involves the estimation of an index of effect magnitude from a series of studies. Typically an estimate such as a standardized mean difference is obtained from each study and these estimates are then combined across studies (Glass, 1978). Statistical procedures have been described for estimating an effect size (standardized mean difference) from a single study (Hedges, 1981; Kraemer & Andrews, 1982) and for combining effect sizes across studies (Hedges, 1982a; Rosenthal & Rubin, 1982). Procedures for fitting models to effect sizes and testing the goodness of fit of those models also have been developed (Hedges 1982b, 1982c, 1983; Hedges & Olkin, 1983).

All the statistical theory for estimators of effect size that has been developed to date relies on the assumption that available effect size estimates are a random (or representative) sample of an unrestricted population of all possible effect size estimates. This is essentially the same assumption that is made in primary data analysis, where the subjects in studies are assumed to be a random sample from some population of interest. The seemingly innocuous

assumptions that effect size estimates can be considered a random sample from a simple, unrestricted population of effect size estimates often is debatable and sometimes is demonstrably false. If the assumption of random sampling of effect size estimates is not met, statistical procedures relying on this assumption may produce misleading results.

This paper examines the problem of bias induced by the censoring of effect size estimates corresponding to statistically insignificant mean differences and explores some potential solutions to the problem of estimating effect size under this form of nonrandom sampling. Some sources of the tendency to censor the results of studies yielding statistically insignificant mean differences are reviewed first. An explicit statistical model that includes restriction of observable data to studies yielding statistically significant mean differences is proposed. The distribution of the sample standardized mean difference is studied under restriction to significant results. Then, the maximum likelihood estimator of the effect size is derived under the model in which only significant results are observed. The bias of this maximum likelihood estimator is studied, and the exact distribution of the estimator is obtained. Methods for estimation of effect size from a series of experiments also are considered. Finally, some of the methods presented in this paper are applied to some effect size data.

The Problem of Nonrandom Sampling of Effect Size Estimates

Failure To Report Statistics for Nonsignificant Results

One readily demonstrable source of bias in reported estimates of effect size stems from the failure to report the details of the outcomes of statistical analyses when mean differences are not statistically significant. Calculation of an effect size estimate requires the values of sample means and standard deviations or the value of a test statistic from a t or F test for the difference between means. Authors sometimes fail to report either the test statistics or descriptive statistics but simply report "no significant difference." This tendency to report the details of statistical analyses only when significant results are obtained is part of what has been labeled "prejudice against the null hypothesis." Although statisticians are unlikely to condone such conditional reporting of statistical results, the practice is prevalent in educational and psychological research (see Greenwald, 1975). Apparently, statistical indoctrination about null hypothesis tests has led many researchers to believe that it is incorrect to interpret not only hypothesis tests but the data as a whole when the null hypothesis cannot be rejected. One consequence of this practice is that the data needed to calculate effect size estimates are sometimes selectively omitted in research reports when insignificant results are obtained. For example, as much as 40% of the potential effect size estimates in two recent

meta-analyses (Eagly & Carli, 1981; Strube, 1981) could not be calculated because of incomplete reporting of nonsignificant results.

Failure To Publish Nonsignificant Results

Another potential source of bias in estimates of effect size stems from the purported tendency of journal editors not to publish research studies that fail to produce statistically significant results. Sterling (1959) argued that nonsignificant results are rarely published and therefore the published literature may be full of Type I errors. A survey of three psychology journals by Sterling (1959) and a more extensive survey of psychology journals by Bozarth and Roberts (1972) showed that 97% and 94%, respectively, of the articles examined that used statistics rejected the statistical null hypothesis at the $\alpha = .05$ significance level. If the studies examined by Sterling and by Bozarth and Roberts are representative of all studies conducted in psychology, their results imply either a large number of Type I errors, an average power in excess of .90, or some combination of these possibilities. The overwhelming proportion of articles rejecting the null hypothesis in these surveys is indeed remarkable, given the results of studies of the power of statistical tests in psychological research (Chase & Chase, 1976; Cohen, 1962). These surveys of statistical power suggest that the average power of statistical tests in psychological research is between .25 and .85, depending on the assumed magnitude of effects. Therefore, the surveys of statistical power suggest that between 25% and 85% of the studies in psychology journals would be expected to yield statistically significant results.

Thus, there is at least some reason to believe that journal articles contain a greater than expected proportion of statistically significant results and relatively fewer nonsignificant results than would be expected from a random sample of *all* studies actually conducted. It is important to recognize that the empirical data are not overwhelming and that surveys of statistical power involve assumptions (e.g., about effect size) that may not be tenable. Moreover, data from the surveys of statistical power involved a somewhat different journal data base than did the studies of Sterling (1959) and of Bozarth and Roberts (1972).

Other sources provide direct evidence of an *intent* of editorial policy to discourage publication of research that failed to yield statistically significant results (see Bakan, 1966; Sidman, 1960). For example, an editorial by Melton (1962) explicitly included statistical significance as one of the most important criteria used to select manuscripts for the *Journal of Experimental Psychology*. A more recent survey of reviewers for major journals in psychology suggests that statistical significance remains an important criterion in reviewing manuscripts for publication (Greenwald, 1975).

Note that the *perception* of an editorial policy that uses statistical significance as a criterion for publishing manuscripts may have nearly the same effect as the actual policy. If authors believe that manuscripts lacking statistically significant results will not be published, they are unlikely to submit such manuscripts for publication.

Consequences of Nonrandom Sampling of Effect Size Estimates

If effect size estimates corresponding to statistically insignificant results are less likely to be sampled (or to be available for sampling), then the sample of effect size estimates will be biased. Because the test statistic for the t test is montonically related to the effect size, studies that produce statistically significant results tend to have effect sizes that are larger in absolute magnitude. Thus, for positive population effect sizes, overrepresentation of studies producing significant differences will tend to bias the sample toward larger effect sizes. Hence each effect size estimate or average of estimates from such a sample will overestimate the absolute magnitude of the population effect size.

Lane and Dunlap (1978) studied an extreme form of bias toward significant results: the situation in which only experiments yielding statistically significant results are observed. They simulated the results of a large number of two-group experiments and selected for further study experiments that yielded statistically significant mean differences. As expected, they found that the mean difference estimated from experiments yielding significant results overestimated the population mean difference. Similarly, they found that estimates of Hays' magnitude index ω^2 were also upwardly biased.

Notation and Model

Suppose that the data arise from a two-group experiment, and let Y_j^E and Y_j^C be the j^{th} experimental (E) and control (C) group scores in the experiment. Assume that Y_j^E and Y_j^C are independently normally distributed within the groups of the experiment, that is,

$$Y_j^E \sim N(\mu^E, \sigma^2),\ j = 1, \ldots, n,$$
$$Y_j^C \sim N(\mu^C, \sigma^2),\ j = 1, \ldots, n.$$

The *effect size* for the experiment is the parameter

$$\delta = \frac{\mu^E - \mu^C}{\sigma}. \tag{1}$$

The effect size is the population value of the treatment effect (mean difference) if the dependent variable is scaled to have unit variance within groups. Note that the effect size is invariant under linear transformations of the outcome variable.

Estimators of Effect Size

Define the estimator g of δ via

$$g = \frac{\overline{Y}^E - \overline{Y}^C}{S}, \tag{2}$$

where $\overline{Y}^E$ and $\overline{Y}^C$ are the experimental and control group sample means and S is the pooled within-group sample standard deviation. Hedges (1981) has shown that as $n \to \infty$, g has an asymptotic distribution given by

$$\sqrt{n}(g - \delta) \sim N[0, \sigma^2(\delta)], \tag{3}$$

where $\sigma^2(\delta)$ is given by

$$\sigma^2(\delta) = 2(1 + \delta^2/8). \tag{4}$$

Thus g is a consistent estimator of δ. Hedges also showed that the bias of g is less than 5% if $n \geq 10$ and obtained a simple unbiased estimator of δ.

A Model for Restriction to Significant Results

Suppose that the data are only observed (e.g., the data are only reported) if the mean difference is statistically significant at some preset significance level α. Then we do not observe the data $Y_1^E, \ldots Y_n^E$ and $Y_1^C, \ldots, Y_n^C$ unless the F test for the difference between means is significant, that is, when,

$$\frac{n(\overline{Y}^E - \overline{Y}^C)^2}{2S^2} > F(\alpha, n), \tag{5}$$

where $F(\alpha, n)$ is the $100(1 - \alpha)$ % critical value of the F distribution with 1 and $(2n - 2)$ degrees of freedom. In the development that follows, we will only need to deal with sample data through the sample means and the pooled within-groups variance (jointly sufficient statistics).

We denote the *observed* experimental and control group sample means by $(\overline{Y}_*^E)$, $(\overline{Y}_*^C)$ and the *observed* difference between the experimental and control group means by $Y_* = (\overline{Y}_*^E) - (\overline{Y}_*^C)$.

Similarly, we denote the *observed* pooled within-group sample variance by S_*^2. Note that the distributions of the *observed* sample means and variance are not the same as the distributions of the means and variances of all studies (including those that cannot be observed in the present model). This is most evident when the joint distribution of the observed mean difference Y_* and the observed variance S_*^2 are considered. All values of (Y_*, S_*^2) with $S_*^2 > n\overline{Y}_*^2 / 2F(\alpha, n)$ occur with zero probability. The same values of the statistics $\overline{Y} = \overline{Y}^E - \overline{Y}^C$ and S^2 (the mean difference and variance when all studies may be observed) occur with positive probability. The restriction to significant mean differences results in a tendency to observe larger absolute mean differences relative to the variance. Geometrically, the restriction to

significant results is equivalent to restricting the observed (Y_*, S_*^2) values to lie between the parabola $S_*^2 = \overline{Y}_*^2 (n/2F(\alpha, n))$ and the $\overline{Y}_*$ axis.

We denote the observed effect size estimate by

$$g_* = \frac{\overline{Y}_*^E - \overline{Y}_*^C}{S_*} = \frac{\overline{Y}_*}{S_*}.$$

Note that the restriction to significant mean differences implies that values of g_* smaller in absolute magnitude than $\sqrt{2F(\alpha, n)/n}$ can never be observed.

Limitation. The model proposed in this paper for restriction to significant results involves strict censoring of all results that are not significant at the $\alpha = .05$ level. More complicated censoring schemes are possible. In many practical situations the censoring rule will not be known. Hence, it may be impossible to tell whether the model described in this section holds, and consequently whether the results described in subsequent sections of this paper apply. In other situations the censoring rule may be known. For example, if the censoring is a result of failure to report statistics for mean differences that are not statistically significant at the $\alpha = .05$ level, then the model described in this section is appropriate. It is important to remember that the results described in this paper depend on the particular model described in this section. Application of the methods presented in this paper may be misleading if the model described in this section is not appropriate.

Distribution of the Effect Size Under Restriction to Significant Results

The simulations conducted by Lane and Dunlap (1978) confirmed some intuitions about the bias of estimators derived from the restricted population of studies yielding significant results. They showed that $\overline{Y}_*$ overestimates $(\mu^E - \mu^C)$ when $\mu^E - \mu^C > 0$. This result is sensible because "small" mean differences (leading to small F statistics) have a smaller probability of being observed than do "large" mean differences. Similarly, they obtained results that suggested that S_* underestimates σ. This, too, is a sensible result because "large" variances (leading to small F statistics) have a smaller probability of being observed. Finally, they observed that estimates of Hays' ω^2, which are derived from g_*, tend to overestimate ω^2 when $\mu^E - \mu^C > 0$. Again, this is a sensible result because "small" values of g_* cannot be observed.

We obtain the distribution of g_* by recognizing that g_* is simply the effect size estimator g subject to censoring of all values between $-\sqrt{2F(\alpha, n)/n}$ and $\sqrt{2F(\alpha, n)/n}$. Thus, the probability density function $h_*(x \mid \delta, n)$ of g_* is simply

$$h_*(x \mid \delta, n) = \begin{cases} \dfrac{h(x \mid \delta, n)}{A(\delta, n, \alpha)} & \text{if } x^2 > 2F(\alpha, n)/n, \\ 0 & \text{if } x^2 \le 2F(\alpha, n)/n, \end{cases} \tag{6}$$

where $h(x|\delta, n)$ is the probability density function of $1/\sqrt{n/2}$ times a noncentral t variate with $(2n-2)$ degrees of freedom and noncentrality parameter $\delta\sqrt{n/2}$, and $A(\delta, n, \alpha)$ is the probability that a noncentral F variate with 1 and $(2n-2)$ degrees of freedom and noncentrality parameter $n\delta^2/2$ exceeds the critical value $F(\alpha, n)$. Formulas and the details of calculating $h(x|\alpha, n)$ are discussed in the Appendix.

The moments of g_* are easily obtained by numerical integration. The expected value and variance of g_* are given in Table I for $\delta = .25, .50, .75, 1.00$, and 1.5, and $n = 10, 20, 30, 40, 50$. As expected, the bias of g_* (as an estimator of δ) tends to be large for small n and small (but nonzero) values of δ. The bias decreases rapidly as δ increases so that the bias is less than 10% for $\delta > 1.0$ and $n \geq 20$ per group. For smaller δ, the bias may be quite severe. If $\delta = .5$ and $n = 15$, the bias in g_* is nearly 100%. If $\delta = .25$, the bias of g_* is over 200% even if $n = 40$ per group. The absolute bias $[E(g_*) - \delta]$ tends to zero as $\delta \to 0$, and tends to zero for large δ. The absolute bias has a maximum between $\delta = .3$ (for $n = 10$) and $\delta = .2$ (for $n = 40$). Thus, the absolute bias is largest for small but nonzero effect sizes. The relative bias $[E(g_*)/\delta]$ increases as $\delta \to 0$ but tends to one as $|\delta|$ becomes large. The bias of g_* as a function of δ is illustrated in Figure 1.

Unfortunately, many quantitative research syntheses are likely to involve studies with sample sizes of less than 40 per group and effect sizes less than 1.00. Hence, the severe bias of g_* for small n and small δ may lead to substan-

TABLE I

Expected Value and Variance of the Sample Effect Size g_* When Only Significant Results ($\alpha = .05$) May Be Observed

	$\delta = 0.25$		$\delta = .50$		$\delta = .75$		$\delta = 1.00$		$\delta = 1.50$	
n	$E(g_*)$	$V(g_*)$	$E(g_*)$	$V(g_*)$	$E(g_*)$	$V(g_*)$	$E(g_*)$	$V(g_*)$	$E(g_*)$	$V(g_*)$
10 EXACT	1.01	.44	1.22	.11	1.30	.10	1.39	.13	1.67	.24
SIMULATED	.99	.48	1.23	.11	—	—	1.39	.14	1.67	.24
20 EXACT	.77	.08	.87	.04	.96	.06	1.10	.08	1.53	.14
SIMULATED	.77	.08	.87	.04	—	—	1.09	.08	1.54	.14
30 EXACT	.65	.03	.73	.03	.85	.05	1.03	.07	1.52	.09
SIMULATED	.65	.03	.73	.03	—	—	1.03	.07	1.53	.09
40 EXACT	.57	.02	.65	.02	.80	.04	1.01	.06	1.52	.07
SIMULATED	.57	.02	.65	.02	—	—	1.01	.06	1.52	.07
50 EXACT	.51	.01	.60	.02	.77	.04	1.00	.05	1.50	.05
SIMULATED	.52	.01	.60	.02	—	—	1.00	.05	1.51	.06

Note. Exact results were obtained by numerical integration. Simulated results were obtained from 2,000–10,000 replications for each combination of sample size and effect size (see Appendix).

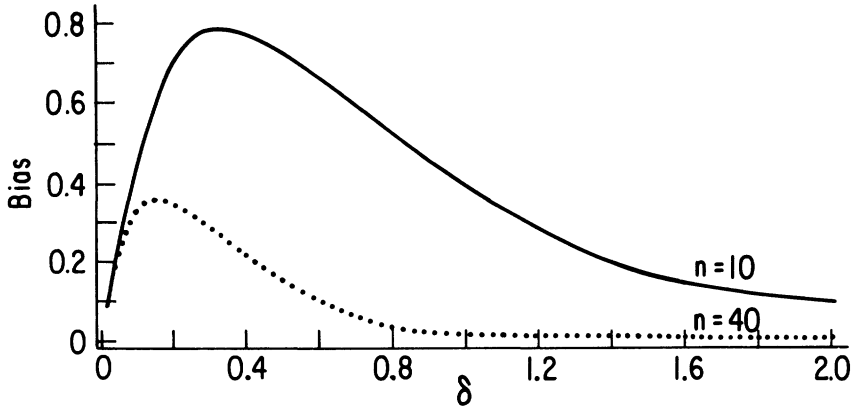

FIGURE 1. The bias of the observed effect size g_* as a function of the population effect size δ when only effect sizes corresponding to mean differences that are statistically significant at the $\alpha = .05$ level may be observed.

tial overestimates of effect size if nonsignificant results are not reported. This finding parallels the results obtained by Lane and Dunlap (1978), who found a severe bias in estimators of a different index of effect magnitude (Hays' ω^2) when nonsignificant results were not available.

Note that the bias of the observed effect size (or mean difference) has implications for the interpretation of research results from a single study. The results in this section show that journal editorial policies requiring statistical significance for publication have the effect of inflating the values of treatment effects that appear in print. Similarly, the tendency to report and discuss only significant effects in a research study inflates the size of experimental effects that are reported in the literature. Such inflation of observed effects may seriously distort our perception of the magnitude of experimental effects.

Estimation of Effect Size from a Single Study When Only Significant Results Are Observed

The results reported in the previous section suggest that the use of g_* (the sample estimate of effect when only statistically significant outcomes are observed) as an estimator of δ may lead to serious bias. When sample sizes and effect sizes are moderate, the bias of g_* may be substantial enough to influence substantive conclusions. For example, when $n = 15$ per group, and the true effect size $\delta = .50$, the bias of g_* is nearly 100%. Such substantial biases could lead to problems in the interpretation of analyses based on effect sizes.

In this section I consider methods for estimation of effect size from a single study when only statistically significant mean differences may be observed.

The methods I propose involve maximum likelihood estimation of δ based on the distribution of g_*. This procedure implicitly (through the distribution of g_*) corrects for the censoring of g_* values corresponding to statistically insignificant mean differences. First, I derive the maximum likelihood estimator of δ based on a g_* value from a single experiment. I obtain the exact distribution of the maximum likelihood estimator numerically and use this distribution to study the bias of the estimator.

Estimation of Effect Size from g_*

The maximum likelihood estimator $\hat{\delta}$ of the effect size δ is obtained by maximizing the log likelihood of g_* as a function of δ. This corresponds to maximizing

$$\log[h_*(g_*|\delta,n)] = \log[h(g_*|\delta,n)] - \log[A(\delta,n,\alpha)], \tag{7}$$

where $h_*(x|\delta,n)$, $h(x|\delta,n)$, and $A(\delta,n,\alpha)$ are defined as in Equation 6. Because $A(\delta,n,\alpha)$ is the probability of a statistically significant mean difference, $A(\delta,n,\alpha)$ is an increasing function of δ for nonnegative δ, which tends to one as $\delta \to \infty$. Hence, the value of δ that gives the maximum of $\log[h_*(x|\delta,n)]$ will tend to be smaller in absolute magnitude than the value of δ that gives the maximum of $\log[h(x|\delta,n)]$. Thus the same observed value of the standardized mean difference leads to a smaller maximum likelihood estimate of δ in the model where only significant results are observed than in the model where all results are observed. When all results are observed, the maximum likelihood estimate of δ is approximately the observed value of the standardized mean difference. Therefore, we would expect that the maximum likelihood estimator of δ based on g_* is smaller than g_*, especially when g_* is itself reasonably small.

The maximum likelihood estimator $\hat{\delta}$ of δ based on g_* cannot be obtained in closed form, but the likelihood (7) can easily be maximized numerically. Figure 2 is a graphical representation of $\hat{\delta}$ as a function of g_* for $n = 10$, 20, and 40. The observed value of g_* is plotted on the abscissa, and the maximum likelihood estimate $\hat{\delta}$ is plotted on the ordinate. In each case the reference line corresponding to $g_* = \hat{\delta}$ also appears. Only positive values of g_* are plotted because of symmetry. That is, if g_* corresponds to the estimate $\hat{\delta}_o$, then $-g_*$ corresponds to the estimate $-\hat{\delta}_o$.

As expected, small values of g_* lead to much smaller values of δ. This result is illustrated in Figure 2, where the function relating g_* to $\hat{\delta}$ lies below the reference line for small g_*. One interpretation is that barely significant values of g_* tend to be associated with small values of δ; thus, the maximum likelihood estimator is small. Large values of g_* lead to values of $\hat{\delta}$ that are almost identical to g_*. This result is illustrated in Figure 2 where the function relating g_* to $\hat{\delta}$ does not deviate greatly from the reference line for large values of g_*.

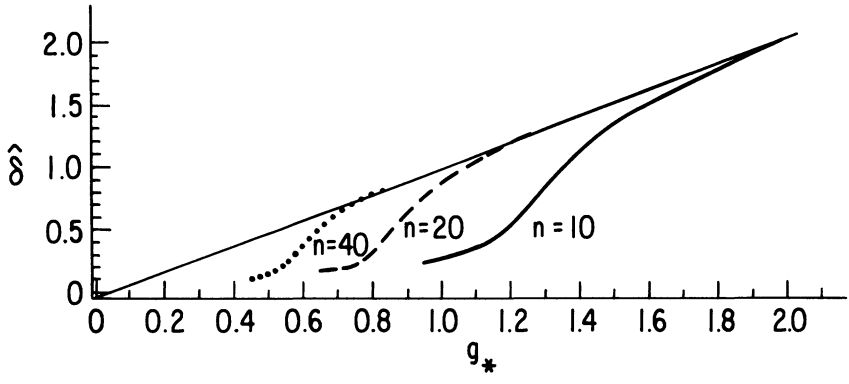

FIGURE 2. The maximum likelihood estimator $\hat{\delta}$ of effect size as a function of the observed effect size g_*.

One interpretation is that large values of g_* arise because of large values of δ and censoring rarely occurs, so that g_* is a good estimate of δ. Similar curves for other sample sizes were generated and they exhibit the same features as the curves given in Figure 2.

One additional feature of the curves in Figure 2 is important. Note that $\hat{\delta}$ decreases as g_* tends to the minimum observable value, but the minimum $\hat{\delta}$ (for positive g_*) is not zero. When $n = 40$, this minimum value is $\hat{\delta} = .111$, whereas for $n = 10$ this minimum value is $\hat{\delta} = .242$. This result does not necessarily imply that these estimates are biased because small population effect sizes are associated with comparable probabilities of positive and negative values of g_*. An effect size near zero will lead to about equal numbers of positive and negative values of $\hat{\delta}$, which may average to numbers smaller than the minimum value of $\hat{\delta}$.

Although the maximum likelihood estimator $\hat{\delta}$ based on g_* is not difficult to obtain numerically, the computations require a specialized computer program. Therefore, we have computed Table II giving $\hat{\delta}$ as a function of g_* for values of n between 10 and 100. Table II gives the maximum likelihood estimator of δ for each value of g_* as well as the minimum observable value of g_* (at the $\alpha = .05$ significance level) and the minimum value of $\hat{\delta}$ for a given n. Table II facilitates calculation of $\hat{\delta}$ without extensive computation by enabling interpolation between tabled values.

Exact Distribution of the Maximum Likelihood Estimator

The maximum likelihood estimator $\hat{\delta}$ of δ is an increasing function of g_*. Because we know the distribution of g_*, we can obtain the distribution of $\hat{\delta}$. Let r denote the function taking $\hat{\delta}$ to g_*. That is, $g_* = r(\hat{\delta})$. Therefore the probability density function $f(x|\delta, n)$ of $\hat{\delta}$ is given by

$$f(x \mid \delta, n) = \left| \frac{dr}{dx} \right| h_*[r(x) \mid \delta, n], \tag{8}$$

where $h_*(y \mid \delta, n)$ is the probability density function of g_* given in Equation 6. The practical problem in using this result is that the explicit mathematical form of the function r is not known. One approach to this problem is to use some smooth approximation to r that can be used for computations. Cubic spline approximations (Ahlberg, Nilson, & Walsh, 1967) were found to give useful approximations to the density function of $\hat{\delta}$.

Plots of the probability density function of $\hat{\delta}$ for $n = 10$ and 40 and $\delta = .0$, .50, and 1.00 are given in Figure 3. Note that the distributions are essentially bimodal, with one peak near the smallest possible estimate and another peak

FIGURE 3. Probability density function of the maximum likelihood estimator $\hat{\delta}$ of effect size.

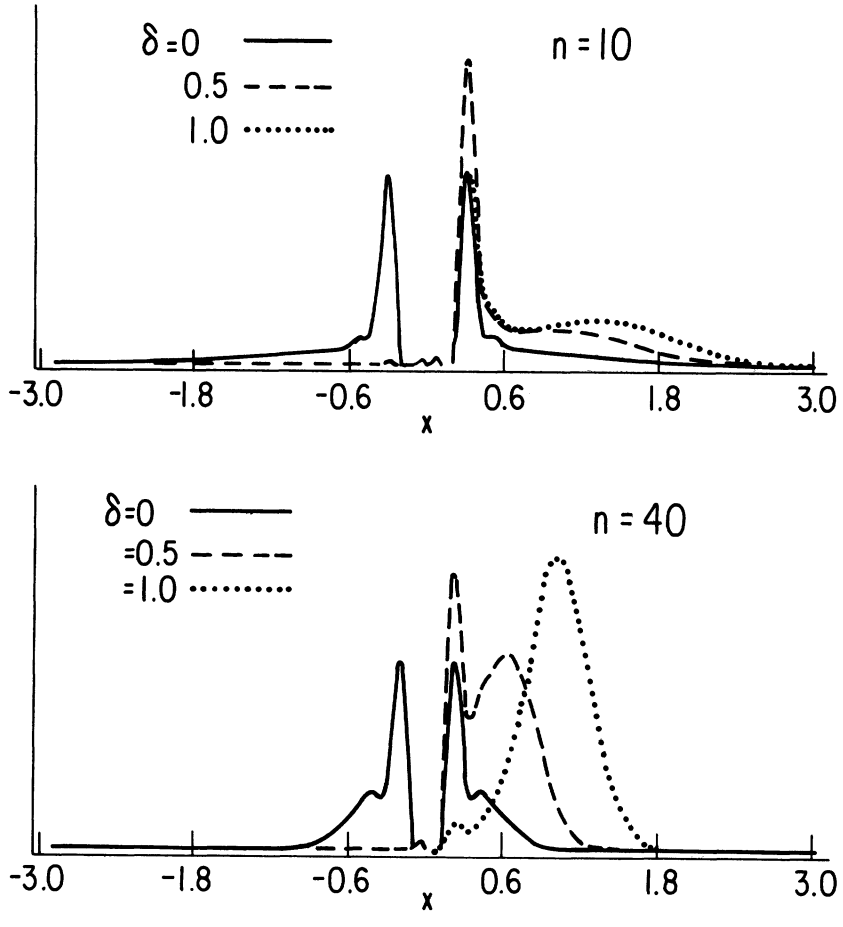

TABLE II

Maximum Likelihood Estimator $\hat{\delta}$ as a Function of Sample Size per Group n and Observed Effect Size g^*

	n																			
	10	12	14	16	18	20	25	30	35	40	45	50	60	70	80	90	100
g_{min}	.940	.847	.777	.722	.677	.640	.569	.517	.477	.445	.419	.397	.362	.334	.312	.294	.279
$\hat{\delta}_{min}$	.242	.216	.197	.183	.171	.161	.142	.129	.120	.111	.108	.101	.092	.085	.079	.074	.070
g^*																	
.30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	.075	.080
.35	—	—	—	—	—	—	—	—	—	—	—	—	—	.090	.100	.116	.145
.40	—	—	—	—	—	—	—	—	—	—	—	.100	.112	.135	.186	.253	.299
.45	—	—	—	—	—	—	—	—	—	.113	.122	.133	.179	.273	.339	.378	.401
.50	—	—	—	—	—	—	—	—	.132	.146	.168	.210	.337	.405	.442	.463	.476
.55	—	—	—	—	—	—	—	.147	.169	.210	.291	.370	.458	.498	.519	.531	.539
.60	—	—	—	—	—	—	.159	.187	.249	.361	.445	.495	.546	.571	.584	.591	.595
.65	—	—	—	—	—	.166	.196	.269	.411	.504	.556	.588	.619	.634	.642	.646	.648
.70	—	—	—	—	.182	.195	.261	.430	.548	.608	.643	.663	.682	.692	.696	.698	.699
.75	—	—	—	.196	.212	.239	.401	.573	.652	.693	.715	.729	.740	.746	.748	.749	.750
.80	—	—	.209	.228	.259	.317	.563	.682	.737	.764	.780	.789	.795	.798	.800	.800	.800
.85	—	.218	.237	.273	.342	.469	.687	.771	.810	.829	.839	.845	.847	.849	.850	.850	.850
.90	—	.243	.279	.349	.498	.631	.788	.847	.875	.888	.895	.898	.899	.900	.900	.900	.900
.95	.246	.277	.340	.494	.659	.757	.872	.916	.935	.945	.948	.950	.949	.950	.950	.950	.950
1.00	.272	.324	.455	.662	.789	.862	.947	.979	.991	.998	1.000	1.001	1.000	1.000	1.000	1.000	1.000
1.05	.305	.394	.626	.801	.896	.951	1.015	1.038	1.047	1.050	1.051	1.052	1.050	1.050	1.050	1.050	1.050

1.10	.347	.523	.782	.916	.987	1.030	1.077	1.094	1.100	1.101	1.102	1.102	1.100	1.100	1.100	1.100	1.100	1.100	1.100	1.100
1.15	.409	.695	.910	1.013	1.068	1.100	1.136	1.148	1.151	1.153	1.153	1.152	1.150	1.150	1.150	1.150	1.150	1.150	1.150	1.150
1.20	.512	.852	1.018	1.098	1.141	1.166	1.193	1.201	1.202	1.203	1.203	1.203	1.200	1.200	1.200	1.200	1.200	1.200	1.200	1.200
1.25	.672	.981	1.111	1.175	1.209	1.228	1.247	1.252	1.253	1.254	1.253	1.253	1.250	1.250	1.250	1.250	1.250	1.250	1.250	1.250
1.30	.844	1.092	1.196	1.246	1.272	1.286	1.300	1.304	1.304	1.303	1.303	1.303	1.300	1.300	1.300	1.300	1.300	1.300	1.300	1.300
1.35	.990	1.189	1.272	1.311	1.332	1.342	1.353	1.354	1.354	1.353	1.353	1.353	1.350	1.350	1.350	1.350	1.350	1.350	1.350	1.350
1.40	1.116	1.276	1.343	1.374	1.389	1.398	1.403	1.404	1.404	1.403	1.403	1.403	1.400	1.400	1.400	1.400	1.400	1.400	1.400	1.400
1.45	1.225	1.355	1.409	1.433	1.445	1.450	1.455	1.454	1.454	1.453	1.454	1.453	1.450	1.450	1.450	1.450	1.450	1.450	1.450	1.450
1.50	1.322	1.429	1.471	1.490	1.499	1.504	1.505	1.505	1.504	1.504	1.504	1.503	1.500	1.500	1.500	1.500	1.500	1.500	1.500	1.500
1.55	1.409	1.497	1.531	1.545	1.552	1.555	1.556	1.555	1.554	1.554	1.553	1.552	1.550	1.550	1.550	1.550	1.550	1.550	1.550	1.550
1.60	1.490	1.562	1.589	1.600	1.604	1.606	1.606	1.605	1.604	1.604	1.603	1.603	1.600	1.600	1.600	1.600	1.600	1.600	1.600	1.600
1.65	1.565	1.622	1.645	1.654	1.656	1.656	1.656	1.655	1.655	1.654	1.654	1.653	1.650	1.650	1.650	1.650	1.650	1.650	1.650	1.650
1.70	1.635	1.683	1.699	1.705	1.707	1.707	1.706	1.706	1.704	1.705	1.704	1.704	1.700	1.700	1.700	1.700	1.700	1.700	1.700	1.700
1.75	1.701	1.741	1.753	1.757	1.758	1.758	1.756	1.755	1.755	1.755	1.754	1.753	1.750	1.750	1.750	1.750	1.750	1.750	1.750	1.750
1.80	1.764	1.796	1.805	1.808	1.808	1.809	1.807	1.806	1.805	1.804	1.804	1.803	1.800	1.800	1.800	1.800	1.800	1.800	1.800	1.800
1.85	1.825	1.850	1.857	1.859	1.859	1.858	1.857	1.855	1.855	1.854	1.854	1.853	1.850	1.850	1.850	1.850	1.850	1.850	1.850	1.850
1.90	1.884	1.905	1.909	1.910	1.910	1.908	1.906	1.905	1.905	1.905	1.904	1.903	1.900	1.900	1.900	1.900	1.900	1.900	1.900	1.900
1.95	1.942	1.956	1.960	1.960	1.959	1.959	1.956	1.956	1.955	1.954	1.954	1.954	1.950	1.950	1.950	1.950	1.950	1.950	1.950	1.950
2.00	1.998	2.010	2.011	2.011	2.009	2.009	2.007	2.006	2.005	2.004	2.004	2.003	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000

Note. The minimum observable g_* value is g_{min} when all observable mean differences are significant at the $\alpha = .05$ level and $\hat{\delta}_{min}$ is the maximum likelihood estimator corresponding to g_{min}.

near the actual value of δ. These distributions are highly nonnormal for δ = .0 and δ = .50. Note also the broad, flat tails of the density function, which suggests that the estimate of δ under censoring is not as precise as the estimator when all studies may be observed.

The density function obtained by cubic spline approximations to $r(x)$ may be integrated numerically to give the expectation and variance of $\hat{\delta}$. The mean and variance of $\hat{\delta}$ for various values of δ are given in Table III. The expected value of $\hat{\delta}$ is not always equal to δ. Therefore $\hat{\delta}$ is a biased estimator and the extent of the bias depends on n and δ. Figure 4 is a plot of the bias $E(\hat{\delta}) - \delta$ as a function of δ for $n = 10$, 20, and 40. Examination of Figure 4 shows that $\hat{\delta}$ overestimates δ for small values of δ. For moderate values of δ the estimator $\hat{\delta}$ underestimates the true value, and $\hat{\delta}$ becomes almost unbiased for large values of δ. In most situations the bias is not substantial. For example, if $n = 20$, the bias never exceeds .19, and if $n = 40$ the maximum bias does not exceed .11. Even when $n = 10$ the bias does not exceed .27.

In many practical situations, the bias may be considerably smaller than the maximum bias. For effect sizes in the range of .5 to .8, the bias never exceeds, .18, even if n is as small as 10. For example if $n = 10$ and δ = .75, the bias is only .06. Similarly, when $n = 20$ and δ = .50, the bias is only .02.

Comparing the expected value of $\hat{\delta}$ with the expected values of g_* given in Table I, we see that the bias of $\hat{\delta}$ is substantially smaller than the bias of g_*. Although $\hat{\delta}$ is not unbiased, $\hat{\delta}$ is much less biased than g_*. Therefore, unless

TABLE III

Mean and Variance of the Maximum Likelihood Estimator $\hat{\delta}$ of Effect Size When Only Significant Results May Be Observed

	δ = 0.25		δ = 0.50		δ = 0.75		δ = 1.00		δ = 1.50	
n	$E(\hat{\delta})$	$V(\hat{\delta})$	$E(\hat{\delta})$	$V(\hat{\delta})$	$E(\hat{\delta})$	$V(\hat{\delta})$	$E(\hat{\delta})$	$V(\hat{\delta})$	$E(\hat{\delta})$	$V(\hat{\delta})$
10 EXACT	.52	.28	.68	.27	.81	.33	.96	.40	1.40	.52
SIMULATED	.51	.30	.70	.27	—	—	.97	.42	1.44	.54
20 EXACT	.39	.10	.52	.13	.69	.18	.92	.21	1.48	.20
SIMULATED	.39	.10	.52	.13	—	—	.92	.21	1.51	.18
30 EXACT	.34	.07	.48	.10	.68	.13	.95	.13	1.49	.11
SIMULATED	.34	.07	.48	.10	—	—	.95	.13	1.53	.10
40 EXACT	.31	.05	.46	.08	.70	.10	.98	.08	1.50	.08
SIMULATED	.31	.05	.46	.08	—	—	.98	.08	1.52	.07
50 EXACT	.28	.04	.44	.07	.71	.07	.99	.06	1.49	.07
SIMULATED	.28	.04	.45	.07	—	—	.99	.06	1.51	.06

Note. Exact results were obtained by numerical integration. Simulated results were obtained from 2,000–10,000 replications for each combination of sample size and effect size (see Appendix).

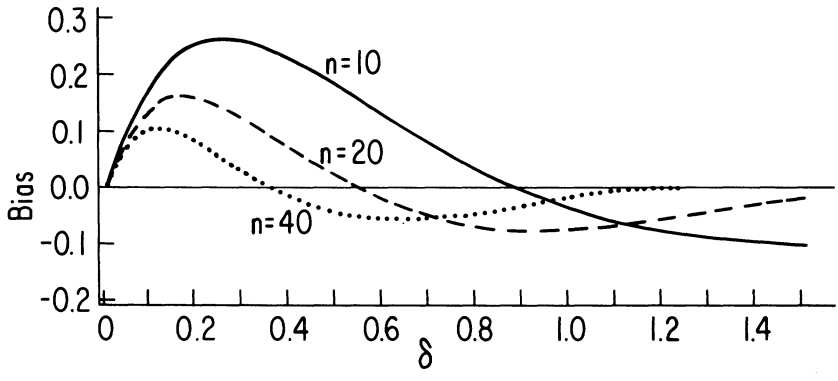

FIGURE 4. Bias $E(\hat{\delta}) - \delta$ of the maximum likelihood estimator $\hat{\delta}$ of effect size.

both sample sizes and effect sizes are quite small, the bias in $\hat{\delta}$ may not be serious enough to affect substantive conclusions in most applications.

It is interesting to compare the efficiency of $\hat{\delta}$ to that of the sample effect size g when the results of all studies may be observed. Using Equation 4 to calculate the variance of g, we see that for $\delta = .25$, the estimator $\hat{\delta}$ has approximately the same variance as does g. When both n and δ are large, the variance of $\hat{\delta}$ is also quite close to the variance of g. For intermediate situations, when δ is moderate to large but n is not large, $\hat{\delta}$ is considerably less efficient than g. For example, when $n = 20$ and $\delta = 1.0$, the variance of $\hat{\delta}$ is nearly twice the variance of g.

Estimation of Effect Size from a Series of Independent Experiments When Only Significant Results Are Observed

In previous sections I examined the problem of estimating the effect size from a single experiment when only statistically significant results are observed. Many practical situations involve the combination of estimates of effect size from several experiments. In this section I discuss methods for estimating effect size from a series of independent experiments when only effect size estimates corresponding to statistically significant mean differences are observed. First, I present a simple counting procedure that may be used to obtain estimates of effect size from a series of studies with homogeneous sample sizes. Next, I discuss maximum likelihood estimation of effect size from a series of experiments. Finally, I suggest the use of a linear combination of estimates derived from each of the individual studies.

Estimation of Effect Size Using Counting Procedures

If the results of a series of identical studies are available, then a simple procedure yields an estimate of the effect size in the unrestricted population

(Hedges & Olkin, 1980). Consider a collection of identical studies, each of which tests the hypothesis that the experimental and control group population means (denoted μ^E *and* μ^C, respectively) are equal. If each experiment has a sample size n per group, then the probability $p^+(\delta, \alpha, n)$ that any given study yields a positive mean difference that is large enough to be statistically significant at the 100α percent level is

$$p^+(\delta, \alpha, n) = \text{Prob}\{X > d_\alpha \mid X \sim t_{2n-2}(\delta\sqrt{n/2})\},$$

where d_α is the 100α percent critical value of the central t distribution, and the notation $X \sim t_m(\lambda)$ means the random variable X has the noncentral t distribution with m degrees of freedom and noncentrality parameter λ. Similarly, the probability $p^-(\delta, \alpha, n)$ that any given study yields a negative mean difference large enough to be statistically significant at the 100α percent level is

$$p^-(\delta, \alpha, n) = \text{Prob}\{X < -d_\alpha \mid X \sim t_{2n-2}(\delta\sqrt{n/2})\}.$$

Given a sample of k studies, each of which has a statistically significant mean difference, the problem is to estimate δ. Each of the k studies can be considered as a Bernoulli trial. The trial is a success if the sample mean difference $(\overline{Y}^E_* - \overline{Y}^C_*)$ is positive. The probability $p(\delta, \alpha, n)$ that any trial is a success is therefore

$$p(\delta, \alpha, n) = \text{Prob}\{X > 0 \mid X > d_\alpha \text{ or } X < -d_\alpha\} = \frac{p^+(\delta, \alpha, n)}{p^+(\delta, \alpha, n) + p^-(\delta, \alpha, n)}.$$

It is well known that the sample number U of successes from a series of k trials has the binomial distribution with parameter $p(\delta, \alpha, n)$.

The maximum likelihood estimator of the parameter $p(\delta, \alpha, n)$ of the binomial distribution on k trials is U/k, the sample proportion of successes. For a given sample size per group n, significance level α, and number of trials k, it is easy to show that $p(\delta, \alpha, n)$ is a strictly monotonic function of δ. Hence, the maximum likelihood estimator of δ, based on the sample number of positive significant results is $\hat{\delta}$, where $\hat{\delta}$ is defined by

$$p(\hat{\delta}, \alpha, n) = \frac{U}{k}.$$

A table for finding $\hat{\delta}$ from values of U/k for various sample sizes n and for the $\alpha = .05$ significance level is given in Hedges and Olkin (1980). A confidence interval for δ can be obtained by first obtaining a confidence interval for $p(\delta, \alpha, n)$. For example, if $[p_1, p_2]$ is a confidence interval for $p(\delta, \alpha, n)$, then $[\delta_1, \delta_2]$ is a confidence interval for δ, where δ_1 and δ_2 are defined by

$$p(\delta_1, \alpha, n) = p_1 \text{ and } p(\delta_2, \alpha, n) = p_2.$$

Confidence intervals for $p(\delta, \alpha, n)$ are obtained by any of the usual methods

of finding a confidence interval for the parameter of the binomial distribution. For large k, the normal approximation to the binomial distribution can be used to obtain asymptotic confidence intervals for $p(\delta, \alpha, n)$. This approximation says that U/k is normally distributed with mean $p(\delta, \alpha, n)$ and variance $p(\delta, \alpha, n)[1 - p(\delta, \alpha, n)]/k$. Details of these procedures are discussed in Hedges and Olkin (1980).

Thus point estimates and a confidence interval for δ in the unrestricted population can be obtained from a sample representing studies that yield a statistically significant mean difference. The method described in this section was applied to some of the results of the simulation reported in Lane and Dunlap (1978). Table IV is a presentation of true parameter values, point estimates, and 90% confidence intervals thus obtained. These results demonstrate the close agreement between estimates obtained from the censored sample and true parameter values.

Counting Estimates When All Significant Mean Differences Are in the Same Direction

In some collections of studies, all statistically significant mean differences may have the same sign. This result is most likely to occur when sample sizes and effect sizes are large, as illustrated by the Lane and Dunlap (1978) simulation. When all significant mean differences have the same sign, the counting method given in the previous section cannot be applied directly, because the

TABLE IV
Estimates and Confidence Intervals for Effect Sizes Based on Counting Procedures

					90% confidence intervals			
n[a]	U[b]	k[c]	δ	$\hat{\delta}$	p_1	p_2	δ_1	δ_2
5	279	332	.25	.26	.8005	.8749	.23	.30
10	300	322	.25	.25	.9045	.9595	.22	.32
15	476	491	.25	.28	.954	.984	.25	.35
20	611	633	.25	.23	.951	.979	.21	.27
5	511	530	.50	.50	.948	.980	.46	.59
10	831	834	.50	.57	.992	.998	.40	.66
15	1313	1315	.50	.50	.996	.999	.46	.57
5	1442	1444	1.00	1.00	.997	.999	.87	1.11

Note. These estimates are based on counts of positive and negative significant results. The values of U, n, and k were derived from data reported in Lane and Dunlap (1978).

[a] Sample size per group.

[b] Number of positive significant ($\alpha = .05$) mean differences.

[c] Number of significant ($\alpha = .05$) mean differences.

maximum likelihood estimate of $p(\delta, \alpha, n)$ is one. In this case, other estimation procedures may be used to obtain an estimate of $p(\delta, \alpha, n)$, which is more intuitively satisfying. It should be noted that the maximum likelihood estimator of δ obtained in the previous section is consistent and asymptotically efficient. When we obtain other estimates of $p(\delta, \alpha, n)$ and transform these estimates to obtain estimates of δ, no invariance theory necessarily guarantees these properties for the estimates of δ. However, the procedure is intuitively sensible.

The situation in which all significant mean differences are in the same direction usually corresponds to a situation in which the true value of $p(\delta, \alpha, n)$ or $1 - p(\delta, \alpha, n)$ is large. This suggests that we may have prior knowledge of the magnitude of $p(\delta, \alpha, n)$. One estimation strategy is to incorporate this prior knowledge into a Bayesian estimate of $p(\delta, \alpha, n)$. For example, we may assume that $p(\delta, \alpha, n)$ has a truncated uniform prior distribution, that is, we assume that there is a p_0 such that any value of $p(\delta, \alpha, n)$ in the interval $[p_0, 1]$ is equally likely.

The Bayes estimate $\hat{p}$ of $p(\delta, \alpha, n)$ based on the truncated uniform prior with k successes in k trials is

$$\hat{p} = \frac{(k+1)\,(1 - p_0^{k+2})}{(k+2)\,(1 - p_0^{k+1})}.$$

For example, if $p_0 = .9$ and there are 1,706 trials, all of which are successes (the Lane and Dunlap result for $n = 20$, $\delta = .5$), the $\hat{p} = .9994$. Transforming $\hat{p}$ to a value of $\hat{\delta}$ gives $\hat{\delta} = .52$. This estimate is quite insensitive to the choice of p_0 for $k = 1{,}706$, but would be more sensitive to p_0 if the number of trials were smaller. A more complete discussion of Bayesian estimation for the parameter of the binomial distribution when all the trials are successes appears in Chew (1971).

Limitations of Counting Estimates

The counting estimates described in this section are simple and intuitively appealing. The counting estimators can sometimes be applied to provide a quick estimate of effect size based on minimal data from a series of studies. These estimators have several limitations that restrict their applicability in practice, however. Counting estimators of effect size requires a fairly large number of studies (counts) to obtain an accurate estimate of $p(\delta, \alpha, n)$. Moreover, the estimate $\hat{\delta}$ of δ obtained by these methods can be sensitive to variations in $\hat{p}\ (\delta, \alpha, n) = U/k$, especially when $\hat{p}(\delta, \alpha, n)$ is close to zero or one. Thus, a large sample of studies is even more important when $p(\delta, \alpha, n)$ is close to zero or one. Formally, the asymptotic theory that underlies vote counting estimators holds as $k \to \infty$. Therefore, vote counting estimators (un-

like other procedures described in this paper) depend on a reasonably large number of studies.

A different limitation stems from the logic of the derivation of the counting estimators. Development in this section depends on the fact that each study has an identical sample size n. Few, if any, collections of research studies consist entirely of studies with identical sample sizes. If the sample sizes do not vary greatly, then a reasonable procedure (Hedges & Olkin, 1980) is to treat the studies as if all had (the same) sample size equal to some average value. The methods in this section can then be applied to obtain an estimate of δ. In many cases, sample sizes of the studies may differ substantially and it may be unreasonable to treat them as equal. It is not obvious how to extend the vote counting methods given in this section to handle unequal sample sizes explicitly. Because unequal sample sizes are the rule rather than the exception in research synthesis, counting estimators are likely to be most useful for providing quick approximate estimates rather than serving as the analytic tool for final analyses.

Maximum Likelihood Estimation of Effect Size

The method of maximum likelihood provides one alternative to simple vote counting methods. Maximum likelihood estimation has the advantage of being applicable to any collection of studies, including those in which the studies do not have identical sample sizes. The method of maximum likelihood also uses more information from each study (e.g., magnitude of the g_* values as well as just the sign) than do vote counting methods. The biggest disadvantage of maximum likelihood estimation is that it requires a specialized computer program to compute the likelihood and its maxima.

Computation of the likelihood of an effect size based on the results of a series of independent experiments is straightforward. Let $\mathbf{g}_* = (g_1, \ldots, g_k)$ be the vector of observed effect sizes from k independent experiments with sample sizes $n_1, \ldots, n_k$ per group, and let δ be the population effect size. Then the log likelihood $L(\delta|\mathbf{g}_*)$ is given by

$$L(\delta|\mathbf{g}_*) = \sum_{i=1}^{k} \log [h(g_i|\delta, n_i)] - \sum_{i=1}^{k} \log [A(\delta, n_i, \alpha)], \tag{9}$$

where $h(x|\delta, n)$ is the probability density function of a noncentral t variate with $(2n-2)$ degrees of freedom and noncentrality parameter $\delta\sqrt{n/2}$, and $A(\delta, n, \alpha)$ is the probability that a noncentral F variate with one and $(2n-2)$ degrees of freedom and noncentrality parameter $n\delta^2/2$ exceeds the $100(1-\alpha)$ percent critical value of the corresponding central F distribution. Details of computing $h(x|\delta, n)$ and $A(\delta, n, \alpha)$ are given in the Appendix.

The maximum likelihood estimate of δ based on the series of experiments is the value $\hat{\delta}$ of δ that maximizes Equation 9. The maximum likelihood

estimate cannot be obtained in closed form, but the estimate can be obtained numerically without much difficulty. Perhaps the simplest procedure is to evaluate Equation 9 for a grid of trial values and then select $\hat{\delta}$ as the value of δ that yields the largest value of $L(\delta|\mathbf{g}_*)$. If a reasonably accurate "first guess" about the value of $\hat{\delta}$ is available (e.g., from a vote-counting estimate or the method given in the next section), the evaluation of a relatively small grid of trial values may give $\hat{\delta}$ to an accuracy of two decimal places or more.

More elaborate numerical procedures could easily be developed. For example, the Newton-Raphson method could be used to iterate rapidly to the value of $\hat{\delta}$. Such procedures would involve the evaluation of fairly complicated derivatives, however, and would probably be warranted only if rather extensive use of the resulting computer programs was envisioned.

Estimation Using a Linear Combination of Estimates Obtained from Individual Experiments

An alternative to simultaneous maximum likelihood estimation of δ based on all k experiments is the linear combination of estimates obtained from each experiment individually. In this procedure, the investigator obtains the maximum likelihood estimates $\hat{\delta}_1, \ldots, \hat{\delta}_k$ from each experiment considered separately. The investigator then obtains a pooled estimate by calculating a linear combination of $\hat{\delta}_1, \ldots, \hat{\delta}_k$. This procedure has several advantages. First, a specialized computer program is not needed because Table II may be used to obtain the maximum likelihood estimates of δ from each experiment. The estimates $\hat{\delta}_1, \ldots, \hat{\delta}_k$ are calculated explicitly and can be examined directly to detect potential outliers or values that seem to deviate greatly from the other estimates. Finally, the analysis using a linear combination of $\hat{\delta}_i$ values corresponds to the analyses usually used in quantitative research syntheses when the results of all studies may be observed (regardless of statistical significance).

The simplest estimator of δ based on $\hat{\delta}_1, \ldots, \hat{\delta}_k$ is the unweighted mean. If the experiments do not all have the same sample size, the unweighted mean will be suboptimal. Intuitively, experiments with large sample sizes will produce better (more precise) estimates of δ and should therefore be given more weight in the combined estimator. Thus a weighted estimator of the form

$$\sum_{i=1}^{k} w_i \hat{\delta}_i / \sum_{i=1}^{k} w_i \tag{10}$$

is more efficient than the unweighted mean. It is easy to demonstrate (see, e.g., Hedges, 1981) that the weights that minimize the variance of (10) are proportional to the reciprocals of the sampling variances of the $\hat{\delta}_i$. That is

$$w_i = \frac{1}{\text{var}(\hat{\delta}_i)}, \; i = 1, \ldots, k,$$

minimize the variance of (10). Unfortunately, the sampling variance of $\hat{\delta}_i$ is a function of δ, the unknown parameter. Thus, the optimal weights cannot be calculated exactly in most applications.

One alternative is to use some weights that are close to optimal but selected in a way that does not involve δ. Because the sampling variance of $\hat{\delta}_i$ is approximately proportional to $1/n_i$, the use of $w_i = n_i$ results in a weighted estimator that is reasonably close to optimal. The weighted estimator also will tend to be less biased than the unweighted mean because estimates from larger experiments (with correspondingly smaller bias) will be given more weight.

Example

Techniques described in this paper were applied to some studies collected by Hedges, Giaconia, and Gage (1981). Data from 10 studies that examined the effects of open and traditional teaching on creativity are given in Table V. Sample size, effect size estimate, and two sample t statistics are given for each study. The difference between the open and traditional group means is statistically significant at the $\alpha = .05$ level in 4 of the 10 studies. The techniques described in this paper are illustrated by their application to the effect size estimates derived from these 4 studies.

The unweighted average of all 10 effect size estimates is .26. The unweighted average of the g_i values obtained in the first 4 studies is .45, somewhat larger than the average for all studies. The unweighted average of the $\hat{\delta}_i$

TABLE V
The Effects of Open vs. Traditional Teaching on Creativity

Study	n_i	g_i	t_i	$\hat{\delta}_i$	$n_i\hat{\delta}_i$	$n_i g_i$
1	90	−.583	−3.91*	−.57	−51.3	−52.47
2	40	.535	2.39*	.21	8.4	21.40
3	36	.779	4.67*	.75	27.0	28.04
4	20	1.052	3.33*	.95	19.0	21.04
5	22	.563	1.87	—	—	12.39
6	10	.308	.69	—	—	3.08
7	10	.081	.18	—	—	.81
8	10	.598	1.34	—	—	5.98
9	39	−.178	−.79	—	—	−6.94
10	50	−.234	−1.17	—	—	−11.70

Note. These data are from Hedges, Giaconia, and Gage (1981).
*$p < .05$.

is .34. In these data, the sample sizes vary considerably, and one might expect that weighted averages would differ from unweighted averages of estimates. The (sample size) weighted average of the effect size estimates for all studies is $g_. = .07$, and the (sample size) weighted average of the $\hat{\delta}_i$ is .02. It is interesting to note that the maximum likelihood estimator of δ based on the first 4 studies is $\hat{\delta} = .01$, which is very close to the weighted average of the $\hat{\delta}_i$.

Applications of the Methods Described in This Article

The statistical procedures described in this paper can be used to provide sharp estimates of effect size under nonrandom sampling when the censoring rule is known precisely. One such application is for the "shrinkage" of effect size estimates that could not have been calculated if the mean difference were not statistically significant. Books or journal articles that report only "not significant" for results that do not attain the $\alpha = .05$ level of significance are an example. In such cases, statistics are reported (and hence an effect size estimate can only be calculated) only if the mean difference is significant. The results of this paper provide a method for correcting such effect size estimates for the effects of censoring nonsignificant results.

The methods presented in this paper also may be applied when the censoring rule is unknown but is believed to be less extreme than complete censoring of all nonsignificant results. For example, some, but not all, nonsignificant results might be censored. The reviewer can impose the more stringent censoring model considered in this paper by using only statistically significant sample effect sizes. The effect size can then be estimated from only the statistically significant sample effect sizes by using the methods described in this paper.

A third application of the methods described in this paper is to examine the potential effects on conclusions of an unknown censoring rule. Here the question is whether a tendency not to sample (publish) studies with statistically insignificant results could have dramatically inflated the observed effect size estimates. The methods described in this paper can be used to obtain an estimate of effect size from the statistically significant sample effect sizes. Comparing this estimate with that obtained by a simple or weighted average of the observed effect sizes (including those that are not statistically significant), the reviewer can evaluate the potential bias due to censoring. If the overall estimates of effect size do not differ greatly, then it is difficult to argue that the overall observed effect size estimate has been inflated by censoring insignificant results without proposing a more extreme and less plausible censoring model. This application of checking the plausibility of censoring as a potential explanation of average effect magnitudes may be one of the most important applications of the methods presented herein.

APPENDIX
Derivations and Computations

Computations for the distribution of g_* were based on $\sqrt{n/2}\, g_*$ and on the corresponding noncentral t distribution with noncentrality parameter $\lambda = \sqrt{n/2}\ \delta$ and $m = 2n - 2$ degrees of freedom. The density function for a noncentral t variate with m degrees of freedom and noncentrality paramenter λ is given (see, e.g., Resnikoff & Lieberman, 1957) by

$$\tilde{h}(x \mid \lambda, m) = \frac{m!}{2^{\frac{m-1}{2}}\Gamma\left(\frac{m}{2}\right)\sqrt{\pi m}} e^{-\frac{1}{2}\frac{m\lambda^2}{m+x^2}} \left(\frac{m}{m+x^2}\right)^{\frac{m+1}{2}} Hh_m\left(\frac{-\lambda x}{\sqrt{m+x^2}}\right),$$

where

$$Hh_m(y) = \int_0^\infty \frac{v^m}{m!} e^{-\frac{1}{2}(v+y)^2}\, dv.$$

For $m < 20$, values of $Hh_m(y)$ can be obtained as

$$Hh_m(y) = P_m(y)\, Hh_0(y) + Q_m(y) Hh_{-1}(y),$$

where $P_m(y)$ and $Q_m(y)$ are polynomials and

$$Hh_0(y) = \frac{1}{2\pi}\int_0^y e^{-\frac{1}{2}t^2}\, dt,$$

and

$$Hh_{-1}(y) = \frac{e^{-\frac{1}{2}y^2}}{2\pi}.$$

A recurrence relationship among the polynomials $P_m(y)$ and $Q_m(y)$ simplifies their computation. For large m it is easier to compute $Hh_m(y)$ by using an asymptotic expansion given by Resnikoff and Lieberman (1957), which is accurate to five decimal places when $m > 20$. This expansion is

$$Hh_m(y) = \frac{1}{m!} t^m e^{-\frac{1}{2}(t+y)^2} \sqrt{\frac{2\pi t^2}{m+t^2}} \left[1 - \frac{3m}{4(m+t^2)^2} + \frac{5m^2}{6(m+t^2)^3}\right],$$

where

$$t = \frac{-y + \sqrt{y^2 + 4m}}{2}.$$

The exact density function of $\hat{\delta}$ was computed by obtaining a table of approximately 40 values of g_* and $\hat{\delta}$ for each n. The tabled values were then used as knot points for cubic spline interpolation between tabled values (e.g., approximation of the function $g_* = r[\hat{\delta}]$) using International Mathematical and Statistical Libraries, Inc. (IMSL) subroutines ICSCCU and ICSEVU. The same splines were used to evaluate the Jacobian $dr/d\hat{\delta}$. Slightly different knot points were used for each value of n to obtain the most accurate approximation, and the final results were checked in several ways, including the simulation study. All of the numerical integrations were computed using IMSL subroutines DCADRE and DCSQDU.

The simulation studies described in this article were conducted using a file of g_i values that were generated for another study (Hedges, 1982a). The g_i values were

derived using standard normal deviates and chi-squared random numbers generated by the IMSL (1977) subroutines GGNML and GGCHS. For each value of δ (.25, .50, 1.00, or 1.50) and for each value of n (10, 20, 30, 40, or 50), 10,000 g_i values were generated using the identity $g = X/\sqrt{S/m}$, where X is a normal variate with variance $2/n$ and mean δ, and S is an independent chi-square variate with $m = 2n - 2$ degrees of freedom. In the present study, values of g_* were obtained by selecting observations with absolute values that exceeded the $\alpha = .05$ critical value of the null distribution of g. Additional values of g were generated when necessary to yield at least 2,000 values of g_* for each combination of n and δ. The g_* values were transformed to $\hat{\delta}_i$ values using cubic spline interpolation on a table of values of g_* and $\hat{\delta}$.

Acknowledgment

This research was supported by the Spencer Foundation.

References

Ahlberg, J., Nilson, E., & Walsh, J. (1967). *The theory of splines and their application.* New York: Academic Press.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 432–437.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27,* 774–775.

Chase, L. J., & Chase, R. B. (1976). Statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61,* 234–237.

Chew, V. (1971). Point estimation of the parameter of the binomial distribution. *American Statistician, 25,* 47–50.

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

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

Glass, G. V (1978). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), *Review of research in education (Vol. 5).* Itasca, IL: Peacock.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

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. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

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

Hedges, L. V. (1982c). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7,* 245–270.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388–395.

Hedges, L. V., Giaconia, R.M., & Gage, N.L. (1981). *The empirical evidence on the effectiveness of open education.* (Final report of the Stanford Research Synthesis Project, Volume II). Stanford, CA: School of Education, Stanford University.

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

Hedges, L. V., & Olkin, I. (1983). Regression models in research synthesis. *American Statistician, 37,* 137–140.

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

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

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

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Resnikoff, G. J., & Lieberman, G. J. (1957). *Tables of the noncentral t-distribution.* Stanford, CA: Stanford University Press.

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

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

Sterling, T. C. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Strube, M. J. (1981). Meta-analysis and cross-cultural comparison: Sex differences in child competitiveness. *Journal of Cross Cultural Psychology, 12,* 3–20.

Author

LARRY V. HEDGES, Assistant Professor, Department of Education, University of Chicago, 5835 S. Kimbark Avenue, Chicago, IL 60637. *Specialization:* Statistics.