Journal of Educational Statistics
Fall 1993, Vol. 18, No. 3, pp. 271–279

Estimation of Effect Size From a Series of Experiments Involving Paired Comparisons

Robert D. Gibbons
Donald R. Hedeker
John M. Davis
University of Illinois at Chicago

Key words: *effect size, paired comparison, meta-analysis*

This article develops the distribution theory for a Glass-type (1976) estimator of effect size from studies involving paired comparisons. We derive an unbiased estimator of effect size following Hedges's (1981) work on combining effect sizes from independent samples. It is shown that studies involving mixtures of paired and independent samples can also be combined in this way. Approximate confidence limits for effect size are derived. The method is illustrated using an example from psychiatric research.

The work of Hedges (1981, 1982, 1983) and Hedges and Olkin (1985) provides a very general statistical approach for combining effect-size-type estimators (e.g., see Glass, 1976) obtained from multiple independent studies. Hedges (1981) showed that the typical effect size estimator (Glass, 1976)—that is, the sample between-group mean difference divided by the sample standard deviation—is a biased estimator of the true population value. Hedges (1981) derived the exact distribution of Glass's estimator and obtained a minimum variance-unbiased estimator of effect size that was shown to have uniformly smaller variance than the Glass-type estimator.

There are, however, situations in which this estimator does not apply. One important case, which is observed frequently in the behavioral sciences, is for within-subject designs in which each subject is repeatedly observed under both treatments or treatment versus control conditions. A further complication that often arises in psychiatric investigations, for example (see Davis, Israni, Janicak, Wang, & Holland, 1990), is that some studies have utilized within-subject designs and others have utilized between-subject designs. In these mixed cases, it may still be of considerable interest to estimate an overall effect size and corresponding confidence limits. For example, Davis et al. (1990) reviewed the nine available studies on the efficacy of the new and potentially promising drug chloripramine for treatment of obsessive-compulsive disease (OCD). Because treatment of this psychiatric disorder with conventional antidepressant drugs has generally had little effect, it was important to review all available double-blind random assignment studies in which chloripramine was compared to standard antidepressants. Unfortunately, five of these studies involved parallel group designs, whereas four involved cross-over designs, making routine application of traditional meta-analytic procedures invalid.

The purpose of this article is to derive distribution theory for the case of related samples and to obtain a corresponding minimum variance-unbiased estimator of effect size. In addition, we explore the possibility of combining effect sizes from studies involving a mixture of related and unrelated samples. The resulting estimators are illustrated using the previously described example from psychiatric research. The results that follow apply equally to cases in which paired or one-sample t statistics would normally be applied on a study-by-study basis.

Structural Model for Related Samples

To begin, we assume that each study is a replication of the others, and, if a common scale of measurement had only been used, the studies could be readily combined, except, perhaps, for a random study specific component of variation. Let D_{ij} represent the jth difference score from the ith experiment. Assume that, for fixed i, D_{ij} has a normal distribution with μ_{Di} and variance σ^2_{Di}. That is,

$$D_{ij} \sim N(\mu_{Di}, \sigma^2_{Di}), \qquad j = 1, \ldots, n_i, \qquad i = 1, \ldots, k, \tag{1}$$

where $\delta_{Di} = \mu_{Di}/\sigma_{Di}$ is termed the effect size, for which we assume $\delta_{Di} = \delta$ for $i = 1, \ldots, k$ (i.e., δ is the mean difference when the response scale has unit variance). Following Hedges (1981), the structural model for the observations can be written as

$$D_{ij} = \sigma_{Di}\,\delta + \epsilon_{ij}, \qquad j = 1, \ldots, n_i, \qquad i = 1, \ldots, k, \tag{2}$$

where $\epsilon_{ij} \sim N(0, \sigma^2_{Di})$. Let $\bar{D}_i$ represent the average difference between the paired observations for study i. Then, for fixed i, $\bar{D}_i$ has a normal distribution with mean μ_{Di} and variance σ^2_{Di}/n_i.

Glass-Type Estimator

For the case in which there are two independent samples per study, Glass (1976) proposed an estimator for δ, based on the sample value of the standardized mean difference for each experiment, which is then averaged over k studies to provide the overall estimator. In the present case, this estimator becomes

$$g_i = \bar{D}_i/S_{Di}, \qquad i = 1, \ldots, k, \tag{3}$$

where

$$\bar{D}_i = \sum_{j=1}^{n_i} D_{ij}/n_i,$$

and

$$S_{Di} = \sqrt{\frac{\sum_{j=1}^{n_i} (D_{ij} - \bar{D}_i)^2}{n_i - 1}} .$$

Few studies, however, directly provide the required information necessary for computing g_i. Typically, these studies will each report the value of a paired t statistic (t_{Di}), the number of paired observations (n_i), and either the mean difference $\bar{D}_i$ or, more typically, the individual condition means $\bar{Y}_{1i}$ and $\bar{Y}_{2i}$. The standard deviation of the paired differences is rarely reported but is clearly required. We note, however, that

$$\bar{D}_i = \bar{Y}_{1i} - \bar{Y}_{2i}$$

and that

$$S_{Di} = \bar{D}_i \sqrt{n_i} / t_{Di} .$$

Hence, the required sufficient statistics can usually be obtained.

The Distribution of g_i

If $\bar{Y}_{Di}$ is distributed as $N(\mu_{Di}, \sigma^2_{Di}/n_i)$, then the sampling distribution of the one-sample or paired t statistic

$$t_{Di} = \frac{\bar{Y}_{Di} - \mu_{Di}}{\sqrt{S^2_{Di}/n_i}}$$

is Student's t with $n - 1$ degrees of freedom. In the case of related samples considered here, $\mu_{Di} = 0$; that is, we are testing the null hypothesis that the population difference is zero. The degrees of freedom of t_{Di} are those associated with S^2_{Di}, which in this case are $n_i - 1$. If $\mu'_{Di} \neq \mu_{Di}$ and if $\bar{D}_i$ is distributed $N(\mu'_{Di}, \sigma^2_{Di}/n_i)$, then the statistic,

$$t' = \frac{\bar{D}_i - \mu_{Di}}{\sqrt{S^2_{Di}/n_i}} = \frac{\bar{D}_i - \mu'_{D_i}}{\sqrt{S^2_{Di}/n_i}} + \frac{\mu'_{Di} - \mu_{Di}}{\sqrt{\sigma^2_{Di}/n_i}} \sqrt{\frac{\sigma^2_{Di}}{S^2_{Di}}} \tag{4}$$

$$= t + \frac{\delta \sqrt{n_i}}{\sqrt{s^2_{Di}/\sigma^2_{Di}}} , \tag{5}$$

where $\delta = (\mu'_{Di} - \mu_{Di})/\sigma_{Di}$, is distributed as noncentral t with $n_i - 1$ degrees of freedom and noncentrality parameter $\delta\sqrt{n_i}$ (e.g., see Bickel & Doksum, 1977, p. 213).

Expectation and Variance of g_i

Hedges (1981) noted that, for the case in which there are two independent samples, g_i is distributed as $1/\sqrt{\tilde{n}_i}$ times a noncentral t variate with $n_{Ci} + n_{Ei} - 2$ degrees of freedom and noncentrality parameter $\sqrt{\tilde{n}_i}\delta$, where $\tilde{n}_i$ is the harmonic mean $n_{Ei}n_{Ci}/(n_{Ei} + n_{Ci})$ of the experimental and control group sample sizes, $\delta = (\mu_{Ei} - \mu_{Ci})/\sigma_i$ and $g_i = (\bar{Y}_{Ei} - \bar{Y}_{Ci})/S_i$. In the one-sample or related sample case, the degrees of freedom are $n_i - 1$, and the noncentrality parameter is $\sqrt{n_i}\delta$, as previously noted. Merrington and Pearson (1958) gave the first moment of $t'\nu$ as

$$\mu_1' = \left(\frac{1}{2}\nu\right)^{1/2} \frac{\Gamma[\frac{1}{2}(\nu-1)]}{\Gamma(\frac{1}{2}\nu)}\phi, \tag{6}$$

where $\Gamma(X)$ is the gamma function (see Mood & Graybill, 1963), ν is the number of degrees of freedom, and ϕ is the noncentrality parameter, which in our case is $\phi = \delta_i\sqrt{n_i} = (\mu_{Di}/\sigma_{Di})\sqrt{n_i}$. Following Hedges (1981), let

$$C(\nu) = \frac{\Gamma(\frac{1}{2}\nu)}{(\frac{1}{2}\nu)^{1/2}\Gamma[\frac{1}{2}(\nu-1)]}; \tag{7}$$

then

$$E(g_i\sqrt{n_i}) = \delta\sqrt{n_i}/C(\nu_i), \tag{8}$$

and the expectation of the Glass-type estimator is

$$E(g_i) = E\left(g_i\sqrt{n_i}\cdot\frac{1}{\sqrt{n_i}}\right) \tag{9}$$

$$= \frac{1}{\sqrt{n_i}}E(g_i\sqrt{n_i}) \tag{10}$$

$$= \delta/C(\nu_i). \tag{11}$$

The variance of $t'\nu$ is given by

$$\mathrm{Var}(t'\nu) = \frac{\nu}{\nu-2}(1+\delta^2) - \mu_1'^2, \tag{12}$$

which, when applied to the current problem, yields:

$$V(g_i) = V\left(g_i\sqrt{n_i}\cdot\frac{1}{\sqrt{n_i}}\right) \tag{13}$$

$$= \frac{1}{n_i}V(g_i\sqrt{n_i}) \tag{14}$$

$$= \frac{1}{n_i}\left[\frac{\nu_i}{\nu_i-2}(1+\delta^2 n_i) - \delta^2 n_i/C^2(\nu_i)\right] \tag{15}$$

$$= \frac{\nu_i}{(\nu_i-2)n_i}(1+\delta^2 n_i) - \delta^2/C^2(\nu_i), \tag{16}$$

which is the one-sample or paired sample equivalent of Equation 6b in Hedges (1981). Of course, in the present context, $\delta = \mu_{Di}/\sigma_{Di}$, and $\nu = n_i - 1$. For computational purposes, Hedges (1981) suggests the approximation,

$$C(\nu) \approx 1 - \frac{3}{4\nu - 1}, \tag{17}$$

which should be adequate for most practical purposes.

An Unbiased Estimator of g_i

The bias of the Glass-type estimator is

$$E(g_i) - \delta = \delta[1 - 1/C(\nu_i)], \tag{18}$$

which follows from (11). The bias depends on the effect size δ and on ν_i, the number of degrees of freedom used in estimating S_{Di}. As such, bias is negligible when n is larger, but it can be appreciable when n is small. This is perhaps a greater problem in the paired sample case since $\nu_i = n_i - 1$ and not $n_{1i} + n_{2i} - 2$, as in the two independent sample case. To remove the bias, we can replace the biased estimator g_i with

$$g_i^u = g_i C(\nu_i). \tag{19}$$

Hedges (1981) has shown that g_i^u is the unique, uniformly minimum variance-unbiased estimator of δ. The proof given by Hedges (1981) extends to the present setting as well.

Testing Homogeneity of Effect Size

Given unbiased effect size estimates from k studies, we must determine whether there is a single effect size δ or k distinct effect sizes (i.e., $\delta_i, i = 1, \ldots, k$). To this end, a homogeneity test analogous to the one derived by Hedges (1982) can be developed in the present context. The large-sample test for equality of k effect sizes is

$$H = \sum_{i=1}^{k} \frac{(g_i^u - g.)^2}{\sigma_i^2(g_i^u)}, \tag{20}$$

where $g.$ is the weighted effect size estimator given in the following section and the approximate large-sample variance is given by

$$\sigma_i^2(g_i^u) = \frac{1}{n_i} + \frac{(g_i^u)^2}{2(n_i - 1)}. \tag{21}$$

If all k studies have the same population effect size, then H has an asymptotic chi-square distribution (i.e., $H \sim \chi^2_{k-1}$). In the event that we reject H_0: $\delta_i = \delta, i = 1, \ldots, k$, then it may be unwise to pool the information across the k studies.

Estimating Effect Size From k Homogeneous Studies

Hedges (1982) and Hedges and Olkin (1985) suggest a weighted estimator of average effect size, so that the most precise effect size estimates (e.g., from the largest studies) are more heavily weighted than the less precise estimates (e.g., from the smaller studies). In the present context, the weighted estimator is given by

$$g. = \left[\sum_{i=1}^{k} \frac{g_i^u}{\sigma_i^2(g_i^u)}\right] \Big/ \left[\sum_{i=1}^{k} \frac{1}{\sigma_i^2(g_i^u)}\right], \tag{22}$$

where, as noted in the previous section, $\sigma_i^2(g_i^u)$ is the large-sample variance of g_i^u. Hedges (1982) suggests that a slightly more precise estimator can be obtained by replacing g_i^u with the provisional value of $g.$ and iterating. Simulation studies indicate that use of the large-sample approximation is reasonable, even when all of the studies have a sample size of only 10 per group.

Confidence Intervals for the Effect Size

To test the null hypothesis (i.e., H_0: $\delta = 0$), we can use the normal approximation to the noncentral t distribution and treat each g_i^u as if it were normally distributed with mean δ and variance $1/n_i + \delta^2/2\nu_i$ (Johnson & Welch, 1939). Thus the distribution of $g.$ is approximated by

$$g. \sim N(\delta, \sigma^2),$$

where

$$\sigma^2 = \left[\sum_{i=1}^{k} 1/\sigma_i^2(g_i^u)\right]^{-1}. \tag{23}$$

Substitution of $g.$ for δ gives $\hat{\sigma}^2$, which can be used to construct an approximate confidence interval for δ—that is,

$$g. - Z_{\alpha/2}\hat{\sigma} \leq \delta \leq g. + Z_{\alpha/2}\hat{\sigma}, \tag{24}$$

where $Z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentage point of the standard normal distribution.

Illustration

The antidepressant drug chloripramine was serendipitously discovered to be effective in the treatment of obsessive-compulsive disease (OCD). Other antidepressant drugs were generally not strikingly effective, so it was hypothesized that chloripramine had unique biochemical properties that are implicated in OCD. There have been nine double-blind random assignment well-controlled studies comparing chloripramine to standard antidepressants. These studies have been of modest size, and five have used parallel group designs (between subjects) while four have used cross-over designs (within subjects). Summary statistics for the meta-analysis of these data are provided in Table 1.

Inspection of Table 1 reveals that all studies showed chloripramine to be more effective than standard antidepressant treatment. The weighted estimate of the overall effect size is $g. = -.7639$ with an estimated variance of

TABLE 1

Summary statistics for the meta-analysis of chloripramine versus a standard antidepressant in the treatment of obsessive compulsive disease

Study	Design[a]	n_1[b]	n_2[c]	Mean[d]	SD[e]	g_i^u[f]	$\sigma^2(g_i^u)$[g]
1	W	13	13	−4.00	4.31	−.86	.11
2	W	10	10	−1.70	2.03	−.76	.13
3	W	12	12	−0.58	1.05	−.51	.09
4	W	48	47	−1.70	2.37	−.70	.03
5	B	16	14	−2.70	3.15	−.81	.27
6	B	16	14	−21.70	17.51	−1.17	.29
7	B	32	30	−1.22	0.98	−1.21	.15
8	B	8	6	−0.27	0.54	−.44	.54
9	B	20	18	−0.40	0.48	−.79	.22

[a] W = within-group, B = between-group
[b] number of chloripramine patients
[c] number of standard treatment patients
[d] mean difference or difference in means
[e] *SD* of differences or pooled *SD*
[f] unbiased effect size estimator
[g] Variance of unbiased effect size estimator

$\sigma^2 = .0226$. The 95% confidence interval for the effect size is $-.7639 - 1.96\sqrt{.0226} = -1.586$ to $-.7639 + 1.96\sqrt{0.226} = -.4692$, which clearly does not include zero. In fact, the one-tailed probability associated with this effect size when expressed as a normal deviate is $-.7639/\sqrt{.0226} = -5.08$, $p < 10^{-7}$. The test for homogeneity of effect sizes yielded $H = 3.00$, $df = 8, p < .93$, which suggests that the results were consistent across the nine studies. More details of the meta-analysis are available in Davis et al. (1990).

The finding of consistent evidence for the superiority of chloripramine from the combined studies substantiates that, of these drugs, chloripramine is the only high potency inhibitor of the neurotransmitter serotonin. It is hypothesized that chloripramine's clinical specificity is due to its biological specificity.

It should be noted that, when combining effect sizes from between- and within-subject studies, greater weight will typically be given to the within-subject studies because pairing or matching will often bring about a reduction in variability. To the extent that the variance estimates are highly dissimilar, this approach to research synthesis may not be useful. In the present illustration, variance estimates from the within-subject studies were smaller than those obtained from the between-subject studies, and the unbiased effect size estimates were more homogeneous. In light of this, giving somewhat greater weight to these studies seems reasonable.

Discussion

The derivation presented in this article provides a generalization of the result of Hedges (1981) for the paired comparison case. The results of this article are exactly as expected—namely, the effect size is now expressed in terms of the ratio of the average difference to the standard deviation of the differences multiplied by a correction factor, and the degrees of freedom are $n_i - 1$. The simplicity of the generalization, however, provides us with the ability to combine evidence from studies involving both between-subject and within-subject designs, by simply substituting the appropriate value of g_i^u into (22), depending on the nature of the comparison in study i (i.e., paired or independent). The results of this article also apply equally to the case of combining one-sample comparisons—for example, testing whether the sample-based average value is equal to some hypothesized population mean value μ across several studies.

The results provided here are also similar to those obtained by Becker (1988), who considered the synthesis of change scores in a different format. Becker (1988) examined the case in which pre- and postmeasures are normally distributed with possibly different means, common variance, and pre- and postmeasure correlation ρ. Estimates of ρ, however, are typically not available in published reports. In addition, for within-subject studies, subjects are often selected to be quite similar at baseline, but they respond to the intervention or treatment quite differently. In light of this, posttreatment variability is often larger than pretreatment variability (i.e., heterogeneous variances). In contrast, we consider the change score and its variance as a single sample problem (i.e., H_0: $\mu_D = 0$); hence the homogeneity of variance assumption and known value of ρ are not required.

References

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics*. Oakland, CA: Holden-Day.

Davis, J. M., Israni, T., Janicak, P., Wang, Z., & Holland, D. (1990). A meta-analysis of drug efficacy studies in obsessive compulsive disorder. *Biological Psychiatry*, *29*, 444. (Abstract No. P-09-20)

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

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

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

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

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

Johnson, N. L., & Welch, B. L. (1939). Applications of the non-central t-distribution. *Biometrika*, *31*, 362–389.

Merrington, M., & Pearson, E. S. (1958). An approximation to the distribution of non-central t. *Biometrika*, *45*, 484–491.

Mood, A. M., & Graybill, F. A. (1963). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Authors

ROBERT D. GIBBONS is Professor of Biostatistics, University of Illinois at Chicago, Biometric Laboratory, 912 S. Wood, Chicago, IL 60612. He specializes in statistical methods.

DONALD R. HEDEKER is Assistant Professor of Biostatistics, University of Illinois at Chicago and Prevention Research Center, 850 W. Jackson, Chicago, IL 60607. He specializes in statistical methods.

JOHN M. DAVIS is Gilman Professor of Psychiatry, University of Illinois at Chicago and Illinois State Psychiatric Institute, Department of Psychiatry, 912 S. Wood, Chicago, IL 60612. He specializes in psychiatric research.