Journal of Educational Statistics
Summer, 1981, Volume 6, Number 2, pp. 107-128

DISTRIBUTION THEORY FOR GLASS'S ESTIMATOR OF EFFECT SIZE AND RELATED ESTIMATORS

Larry V. Hedges

The University of Chicago

Key Words: Meta-analysis, Research synthesis, Standardized mean difference, Measurement error, Weighting of estimators, Invalidity, confidence intervals.

ABSTRACT

Glass's estimator of effect size, the sample mean difference divided by the sample standard deviation, is studied in the context of an explicit statistical model. The exact distribution of Glass's estimator is obtained and the estimator is shown to have a small sample bias. The minimum variance unbiased estimator is obtained and shown to have uniformly smaller variance than Glass's (biased) estimator. Measurement error is shown to attenuate estimates of effect size and a correction is given. The effects of measurement invalidity are discussed. Expressions for weights that yield the most precise weighted estimate of effect size are also derived.

INTRODUCTION

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 are the review of psychotherapy outcome studies (Smith & Glass, 1977) and the review of class size literature (Glass & Smith, 1979). Both examples demonstrate the utility of using quantitative estimates of the magnitude of effects in research reviews.

One class of estimates of effect size is based on the sample mean difference divided by the sample standard deviation. A precise definition of effect size and a structural

model for a series of experiments is given in Section 1. In Section 2 two versions of the standardized mean difference estimator are defined. The distributions of these estimators are given in Section 3. An unbiased version of the estimator that has smaller variance than the biased version is obtained in Section 4. The effect of measurement error on the estimate obtained by standardized mean difference estimators is discussed in Section 5, and a simple correction for measurement error is obtained for response measures of known reliability. The effects of invalidity (unique factors) among response measures are considered in Section 6. Section 7 is a discussion of weighting of estimates from different experiments, and it includes an expression for weights which minimize the variance of the combined estimator. Section 8 gives an example of the application of results presented in this paper.

1. A STRUCTURAL MODEL FOR A SERIES OF EXPERIMENTS

The statistical properties of procedures for combining results from a series of experiments depend on the structural model for the results of the experiments. The structural model used in this paper requires that each experiment uses a response scale from a collection of related measures; that is, each response scale is a linear transformation of a response scale with unit variance within groups. Conceptually, we assume that each experiment to be combined is a replication of the others, differing only in the response scale and sample sizes. That is, the population value of the treatment effect from each study would be identical if transformed to a common response scale. More precisely, let Y^E_{ij} and Y^C_{ij} be the j^{th} scores on the i^{th} experiment from the experimental and control groups, respectively. Assume that for fixed i, Y^E_{ij} and Y^C_{ij} are normally distributed with means μ^E_i and μ^C_i and common variance β^2_i; that is,

$$Y^E_{ij} \sim N(\mu^E_i, \beta^2_i) \quad , \quad j = 1,\ldots,n^E_i \ , \ i = 1,\ldots,k \ ,$$

and

$$Y^C_{ij} \sim N(\mu^C_i, \beta^2_i) \quad , \quad j = 1,\ldots,n^C_i \ , \ i = 1,\ldots,k \ ,$$

where $(\mu^E_i - \mu^C_i)/\beta_i = \delta_i$, $i = 1,\ldots,k$. Assume also that $\beta_i > 0$ and that $\delta_i = \delta$, $i = 1,\ldots,k$. Thus δ, the standardized mean difference, is the treatment effect (mean difference) when the response scale has unit variance. The parameter δ is called the effect size.

A more compact representation of the model presented above involves the explicit use of δ, the within-group standard deviation β_i, a location (scale mean) parameter γ_i, and a residual term ε. This structural model can be written as

$$Y^E_{ij} = \beta_i\delta + \beta_i\gamma_i + \varepsilon^E_{ij} \quad , \quad j = 1,\ldots,n^E_i \quad , \quad i = 1,\ldots,k \, ,$$

$$Y^C_{ij} = \beta_i\gamma_i + \varepsilon^C_{ij} \quad , \quad j = 1,\ldots,n^C_i \quad , \quad i = 1,\ldots,k \, , \qquad (1)$$

where $\varepsilon^E_{ij} \sim N(0,\beta_i^2)$ and $\varepsilon^C_{ij} \sim N(0,\beta_i^2)$.

2. GLASS'S ESTIMATOR OF EFFECT SIZE BASED ON THE STANDARDIZED MEAN DIFFERENCE

Glass (1976) proposed an estimator of δ based on the sample value of the standardized mean difference for each experiment, which is then averaged in a set of experiments to obtain the estimator based on the series of k experiments.

Let $\bar{Y}^E_i$ and $\bar{Y}^C_i$ be the experimental and control group means for the i^{th} experiment and S^C_i is the sample standard deviation of the control group of the i^{th} experiment, and define:

$$g_i = \frac{\bar{Y}^E_i - \bar{Y}^C_i}{S^C_i} \quad , \; i = 1,\ldots,k. \qquad (2)$$

The estimator, G, for the series of experiments is simply the average:

$$G = \frac{1}{k}\sum_{i=1}^{k} g_i \, . \qquad (3)$$

Glass proposed the use of the standard deviation of the control group to standardize the mean difference. His argument was that pooling two variances could lead to different standardized values of the identical mean difference within an experiment where several treatments were compared to a control. The argument depends on the fact that sample standard deviations of each group will surely differ. In many cases, however, the model of equal population variances is reasonable, which suggests that the most precise estimate of the population variance is obtained by pooling. Model (1) includes only two groups, with equal population variances, so

Glass's argument does not apply. Hence a modified Glass estimator can be obtained by using a pooled estimate of the standard deviation.

If g_i' is the modified estimator for the i^{th} experiment, then

$$g_i' = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{S_i} , \tag{4}$$

where S_i is the pooled standard deviation estimate; that is,

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

$\bar{Y}_i^E$, $\bar{Y}_i^C$, $(S_i^E)^2$, $(S_i^C)^2$, n_i^E, and n_i^C are the experimental and control group means, variances, and sample sizes respectively in the i^{th} experiment. The modified estimate G' from the series of k experiments is the average of the g_i':

$$G' = \frac{1}{k} \sum_{i=1}^{k} g_i' . \tag{5}$$

In the next section, the distribution of the estimators g_i and g_i' is shown to depend only on the effect size δ and the sample sizes n_i^E and n_i^C in each experiment. The distributions of the two estimators are identical except for the number of degrees of freedom used to estimate the standard deviation. Therefore the two estimators g_i and g_i' are not treated separately in subsequent sections, but a single estimator is presented and the number of degrees of freedom used to estimate the standard deviation is denoted explicitly by m. Hence the estimator g is the case $m = n^C - 1$ and g' is the case $m = n^E + n^C - 2$.

3. DISTRIBUTION OF GLASS'S ESTIMATE OF EFFECT SIZE AS A STANDARDIZED MEAN DIFFERENCE

The distribution of the estimator, g_i, for the i^{th} experiment can be obtained directly from the noncentral t-distribution. Since the combined estimate G is a linear combination of the estimators g_i, i = 1,...,k, the mean, variance, and bias of G can be obtained easily from the dis-

tribution of the g_i. In particular $g_i\sqrt{n_i^E n_i^C/(n_i^E + n_i^C)}$ is distributed as a noncentral t-variate with noncentrality parameter $\delta\sqrt{n_i^E n_i^C/(n_i^E + n_i^C)}$ and degrees of freedom equal to $n_i^C - 1$ or $n_i^E + n_i^C - 2$, depending on whether the control group or pooled estimate of the standard deviation is used. Thus we obtain the mean and variance of g_i, which are stated in the following theorem.

Theorem 1: Suppose that $Y_{ij}^E \sim N(\mu_i^E,\sigma_i^2)$ and $Y_{ij}^C \sim N(\mu_i^C,\sigma_i^2)$, $i = 1,\ldots,k$ as in model (1) with $\delta = (\mu_i^E - \mu_i^C)/\sigma_i$. Then $g_i = (\bar{Y}_i^E - \bar{Y}_i^C)/S_i$, $i = 1,\ldots,k$ have expectation, variance, bias, and mean squared error given by

$$E(g_i) = \delta/c(m_i), \tag{6a}$$

$$Var(g_i) = \frac{m_i}{(m_i - 2)\tilde{n}_i}[1 + \tilde{n}_i\delta^2] - \delta^2/[c(m_i)]^2, \tag{6b}$$

$$Bias(g_i) = E(g_i) - \delta = \delta[1 - 1/c(m_i)], \tag{6c}$$

$$MSE(g_i) = Var(g_i) + [Bias(g_i)]^2 \tag{6d}$$

$$= \frac{m_i}{(m_i - 2)\tilde{n}_i}[1 + \tilde{n}_i\delta^2] + \delta^2[1 - 2/c(m_i)],$$

where $m_i = n_i^C - 1$ or $n_i^E + n_i^C - 2$ depending on whether S_i is a control group or a pooled standard deviation, $\tilde{n}_i = n_i^E n_i^C/(n_i^E + n_i^C)$, $i = 1,\ldots,k$, and $c(m)$ is defined by

$$c(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}}\,\Gamma\left(\frac{m-1}{2}\right)}. \tag{6e}$$

Proof: The numerator and denominator of g_i are distributed independently as $N\left((\mu_i^E - \mu_i^C),\beta_i^2/\tilde{n}_i\right)$ and $\beta_i\sqrt{\chi^2_{m_i}}/m_i$ where m_i and $\tilde{n}_i$ are defined as above. Therefore, g_i is distributed as $(1/\sqrt{\tilde{n}_i})$ times a noncentral t-variate with m_i degrees of freedom and noncentrality parameter $\sqrt{\tilde{n}_i}\,\delta$. Expressions (6a) and (6b) are obtained directly from the mean and variance of the noncentral t-distribution (see e.g.Johnson & Welch,1939). ||

Corollary 1: If g_i, i = 1,...,k, are the estimators defined as in Theorem 1, and if $n_i^E = n_i^C$, i = 1,...,k, then as $m_i \to \infty$, the distribution of the estimators g_i tends to a normal distribution with mean δ and variance $(2/m_i)(1+\delta^2/4)$.

Proof: This is a direct consequence of the large sample normal approximation to the noncentral t-distribution. ||

Remark on the Consistency of G

One consequence of Corollary 1 is that the estimator g_i is a consistent estimator of δ as $m_i \to \infty$. That is, as m_i increases, the estimate g_i of δ approximates the true value δ with greater precision. The reason is that as $m_i \to \infty$ the mean of g_i tends to δ and the variance of g_i tends to zero. If all the $m_i \to \infty$, i = 1,...,k, the estimator G defined in (3) is also a consistent estimator of δ.

A consequence of Theorem 1 is that the estimator G is not a consistent estimator of δ as $k \to \infty$. That is, even though the number of experiments combined increases, the estimator does not necessarily approximate the true value δ more closely. In fact, the estimates can differ from δ by a considerable amount depending on the sample sizes. To see this, consider the example of a collection of experiments with five subjects per group. The estimator **g** has a bias which results in overestimation of δ by approximately 25 percent when four degrees of freedom are used to estimate β_i. Each estimator g_i has the same bias, therefore G is biased by the same amount as each g_i, i = 1,...,k. As k increases, the bias is unchanged, but the variance of G tends to zero. Thus as the number of studies increases, the estimator G estimates the wrong quantity more precisely.

4. AN UNBIASED ESTIMATOR BASED ON THE STANDARDIZED MEAN DIFFERENCE

As a consequence of Theorem 1, we see that the bias of the estimator g depends only on the effect size δ and on m, the number of degrees of freedom used to estimate the standard deviation in the definition of g. This bias approaches zero when m is large but it can be substantial when m is small. Figure 1 is a graphic representation of the ratio of the expectation of the estimator g to the true value of the effect size parameter δ. Note that for small values of m, this ratio is considerably larger than one, which indicates the estimator g seriously overestimates δ. It therefore seems reasonable to

obtain a modified estimator which is unbiased. Another desirable characteristic of the estimator when corrected for bias is that the bias correction decreases the variance of the estimator.

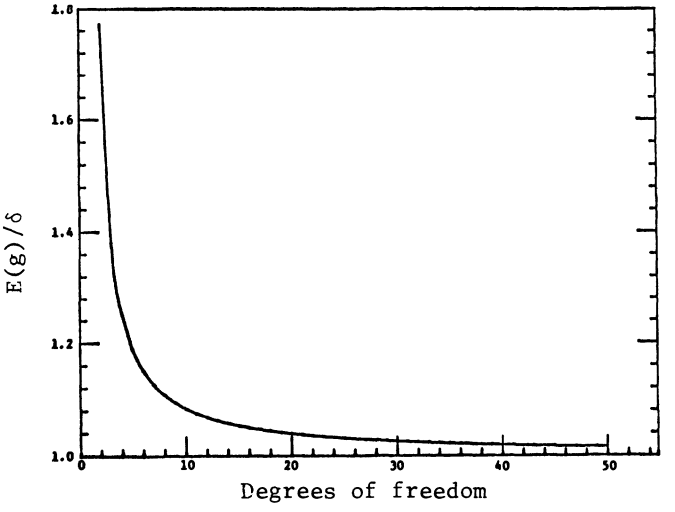

Figure 1

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

The unbiased estimator g_i^U for the i^{th} experiment (using m_i degrees of freedom to estimate β_i) is

$$g_i^U = g_i c(m_i), \quad i = 1,\ldots,k \quad , \tag{7}$$

where c(m) is given by (6e). The unbiased estimator G^U from a series of k experiments is

$$G^U = \frac{1}{k} \sum_{i=1}^{k} g_i^U \quad . \tag{8}$$

The values of the unbiased estimators g_i^U can, therefore, be calculated from the values of the (biased) estimators g_i and a table of values of the correction factor, c(m). Note that since c(m) depends only on m, a single table can be used whether or not sample sizes for experimental and control groups are equal, and whether control group or pooled standard deviations are used in the calculation of g. The values

of c(m) for m = 2(1)50 are tabulated to five decimal places in Table I. The values in this table were obtained using the FORTRAN intrinsic double precision gamma function routine.

An accurate approximation for c(m) can be derived which is satisfactory for most applications. This approximation has the virtue that it can be computed algebraically when using packaged computer programs. The approximation is

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

This approximation has a maximum error of .007 when m = 2, and is accurate to within .00033 when $m \geq 10$. For m greater than 50, the error does not exceed 1.5×10^{-5}.

The correction factor c(m) is always smaller than one, so the variance of g^U is smaller than that of g because

$$\mathrm{Var}(g^U) = \mathrm{Var}[c(m)g] = [c(m)]^2 \, \mathrm{Var}(g) < \mathrm{Var}(g).$$

For small values of m, the variance of g^U will be much smaller than the variance of g. The ratio of the variance of g^U to that of g is plotted as a function of m for $2 \leq m \leq 50$ in Figure 2. Values of $\mathrm{Var}(g^U)/\mathrm{Var}(g) = [c(m)]^2$ are also tabulated in Table I for m = 2(1)50.

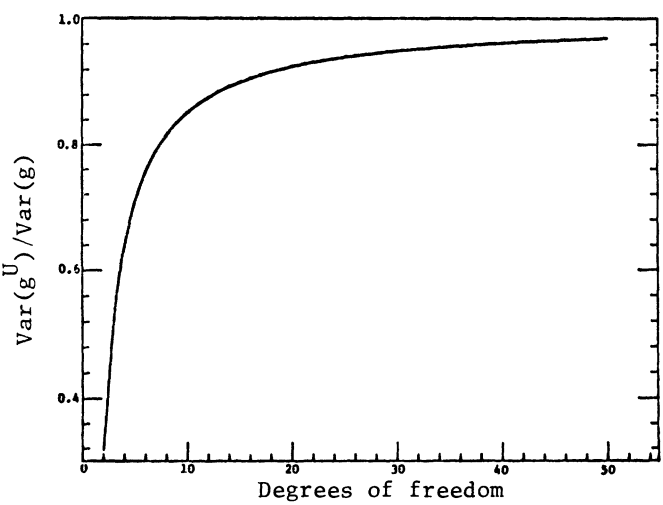

Figure 2

The ratio $\mathrm{Var}(g^U)/\mathrm{Var}(g)$ of the variance of the unbiased standardized mean difference estimator g^U to that of the biased standardized mean difference estimator $g=(\bar{Y}^E - \bar{Y}^C)/S$ as a function of the degrees of freedom of S.

TABLE I

Table of the Ratio $c(m) = g^U/g$ for Obtaining the Unbiased Standardized Mean Difference Estimator g^U from $g = (\bar{Y}^E - \bar{Y}^C)/S$, Where S Uses m Degrees of Freedom to Estimate β.

m	c(m)	$[c(m)]^2$	m	c(m)	$[c(m)]^2$
2	0.56419	0.31831	26	0.97083	0.94250
3	0.72360	0.52360	27	0.97192	0.94463
4	0.79788	0.63662	28	0.97293	0.94660
5	0.84075	0.70686	29	0.97387	0.94843
6	0.86863	0.75451	30	0.97475	0.95015
7	0.88820	0.78890	31	0.97558	0.95175
8	0.90270	0.81487	32	0.97635	0.95325
9	0.91387	0.83517	33	0.97707	0.95467
10	0.92275	0.85146	34	0.97775	0.95600
11	0.92996	0.86483	35	0.97839	0.95725
12	0.93594	0.87599	36	0.97900	0.95843
13	0.94098	0.88545	37	0.97957	0.95955
14	0.94529	0.89357	38	0.98011	0.96062
15	0.94901	0.90062	39	0.98062	0.96162
16	0.95225	0.90679	40	0.98111	0.96258
17	0.95511	0.91224	41	0.98158	0.96349
18	0.95765	0.91709	42	0.98202	0.96436
19	0.95991	0.92143	43	0.98244	0.96519
20	0.96194	0.92534	44	0.98284	0.96598
21	0.96378	0.92888	45	0.98322	0.96673
22	0.96545	0.93209	46	0.98359	0.96745
23	0.96697	0.93504	47	0.98394	0.96814
24	0.96837	0.93773	48	0.98428	0.96881
25	0.96965	0.94021	49	0.98460	0.96944
			50	0.98491	0.97005

These results can be summarized in the following theorem:

Theorem 2: If g_i^U, $i = 1,\ldots,k$ and G^U are the estimators defined by (7) and (8), n_i^E and n_i^C are the experimental and control group sample sizes respectively, m_i degrees of freedom are used to estimate the standard deviation in g_i^U, and the population effect size is δ for each experiment, then the mean and variance of G^U are

$$E(G^U) = \delta,$$

that is, G^U is an unbiased estimate of δ, and

$$\mathrm{Var}(G^U) = \frac{1}{k^2} \sum_{i=1}^{k} \left\{ \frac{m_i[c(m_i)]^2}{(m_i - 2)\tilde{n}_i} (1 + \tilde{n}_i\delta^2) - \delta^2 \right\},$$

where $\tilde{n}_i = n_i^E n_i^C / (n_i^E + n_i^C)$ and $c(m)$ is given by (6e).

When the experimental and control group sample sizes n_i^E and n_i^C are equal, and $m_i = n_i^E + n_i^C - 2$, g_i^U is not only an unbiased estimator of δ, but the unique minimum variance unbiased estimator of δ. This means that no other unbiased estimator can have smaller variance than g_i^U.

Theorem 3: When g_i^U is defined as in (7), $n_i^E = n_i^C$, and $m_i = n_i^E + n_i^C - 2$, then g_i^U is the unique uniformly minimum variance unbiased estimator (UMVUE) of δ.

Proof: If $n_i^E = n_i^C$, then define $X_{ij} = Y_{ij}^E - Y_{ij}^C$, $j = 1,\ldots,n_i^E$, for each $i = 1,\ldots,k$. Each $X_{ij} \sim N(\beta_i\delta, 2\beta_i^2)$, $i = 1,\ldots,k$, and the sample mean and variance of the X_{ij} are jointly complete and sufficient statistics for δ and β_i, $i = 1,\ldots,k$. The estimator $g_i^U = \sqrt{2}c(m)(\bar{X}_i)/\sqrt{V_{X_i}}$ is an unbiased estimator of δ, where $\bar{X}_i$ and V_{X_i} are the sample mean and variance of the X_{ij}, $j = 1,\ldots,n_i^E$. Hence, g_i^U is the unique uniformly minimum variance unbiased estimator of δ by the theorem of Lehmann and Scheffé (1950). ||

4.1 Confidence Intervals for the Effect Size δ

It is reasonable that the estimation of effect size is

the primary goal of combining the results of a series of experiments. Occasionally, a test for the statistical significance or confidence intervals for the effect size may be desired. The exact distribution of a linear combination of central t-variates has recently been obtained by Walker and Saw (1978). Under the null hypothesis H_0: $\delta = 0$, each of the g_i^U has the distribution of a constant multiplied by a central t-variate. Hence, G^U has the distribution of a linear combination of central t-variates under H_0, which can be obtained in closed form from the result of Walker and Saw. When the number of experiments to be combined (k) is large or the degrees of freedom (m_i) of the experiments are large, the computations involved in the method of Walker and Saw become tedious. In this case, the normal approximation to the noncentral t-distribution is quite good (see Johnson & Welch, 1939) and much simpler to compute. The normal approximation to the noncentral t-distribution treats each g_i^U as if it were normally distributed with mean δ and variance $(1/\tilde{n}_i) + \delta^2/2m_i$. Thus the distribution of G^U can be approximated by

$$G^U \sim N(\delta, \sigma^2),$$

where

$$\sigma^2 = \frac{1}{k^2} \sum_{i=1}^{k} \left(\frac{1}{\tilde{n}_i} + \frac{\delta^2}{2m_i} \right).$$

Substituting the consistent estimate G^U for δ in the expression for σ^2 gives $\hat{\sigma}^2$ which can be used to construct an approximate confidence interval for G^U. Specifically, the $100(1-\alpha)$ percent confidence interval for δ is given by

$$G^U - z_{\alpha/2}\hat{\sigma} \leq \delta \leq G^U + z_{\alpha/2}\hat{\sigma},$$

where $z_{\alpha/2}$ is obtained from the standard normal table.

5. THE EFFECT OF ERRORS OF MEASUREMENT

The standardized mean difference δ defined as in (1) is a measure of the magnitude of the treatment effect compared to the variability within the two groups of the experiment. The implicit assumption is that the variability within the experimental and control groups arises from stable differences between subjects (or more generally between experimental units). If the response measure is not perfectly reliable, that is, if errors of measurement are present, then measure-

ment error also contributes to the within-group variability. Measurement error, therefore, alters the population value of the standardized mean difference. If the object is to estimate the value, δ, of the standardized mean difference when no errors of measurement are present, some procedure to correct for measurement error is necessary. In Section 5.1, a structural model including measurement error is presented, and in Section 5.2, the measurement error problem is formulated in terms of classical test theory. The reliability of the response measure is used to obtain an estimator of δ that is not biased by measurement error in Section 5.3.

5.1 A Model for Errors of Measurement

The structural model (1) for experiments to be combined includes a single residual term ε_{ij} for each subject. This residual term includes all sources of deviation from perfect fit to the rest of the structural model. At least two conceptually distinct sources of contribution to this residual term are imaginable. One source might be called subject-treatment interaction, and accounts for the differential response of different subjects to the treatment. Subject-treatment interaction arises because subjects (or experimental units) are not identical, and hence their response to treatment will not be identical even if the treatment is the same for all subjects. Subject-treatment interaction can also be characterized as the disturbing effect of all unmeasured causal variables, that is, all causal variables except treatment.

Another contribution to the residual term arises whenever the measurements of the dependent variable are not infallible. If the dependent variable is measured fallibly, then the difference between the "true" value and the observed value of the variable represents another contribution to the residual term. Note that errors of measurement are conceptually distinct from subject-treatment interactions.

A structural model incorporating both measurement error and subject treatment interaction can be represented as

$$
\begin{aligned}
Y^{E}_{ij} &= \beta_i\delta + \beta_i\gamma_i + (\xi^{E}_{ij} + \eta^{E}_{ij}),\ j = 1,\ldots,n^{E}_{i}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad i = 1,\ldots,k\ , \qquad (9)\\
Y^{C}_{ij} &= \beta_i\gamma_i + (\xi^{C}_{ij} + \eta^{C}_{ij}),\ j = 1,\ldots,n^{C}_{i},\ i = 1,\ldots,k\ ,
\end{aligned}
$$

where ξ_{ij} denotes the interaction of the j^{th} subject with the treatment in the i^{th} experiment and η_{ij} is the error of measurement. If $\eta_{ij} = 0$ for each i and j, model (9) reduces to (1).

Note that the two terms ξ and η always occur together and are therefore indistinguishable without further information. If information on the reliability of the response measure is available, it is possible to determine the variances of ξ and η separately.

5.2 Classical Test Theory for Errors of Measurement

The assumptions of classical test theory for model (9) are that ξ_{ij} and η_{ij}, $i = 1,\ldots,k$, $j = 1,\ldots,n_i$, are independently normally distributed, that ξ_{ij} and η_{ij} are independent of ξ_{st} and η_{st} whenever $i \neq s$ and $j \neq t$, and that both ξ and η have zero expectation within experimental groups. We further assume that the variances of ξ_{ij} and η_{ij} (denoted by $\sigma^2_{\xi_i}$ and $\sigma^2_{\eta_i}$, respectively) are the same for the experimental and control groups of the same experiment. These assumptions allow us to rewrite the within-group variance σ^2_i of the i^{th} experiment as $\sigma^2_i = \sigma^2_{\xi_i} + \sigma^2_{\eta_i}$.

In classical test theory, the structural model is usually simplified into an equation with two terms: true score and error. The structural equation would, therefore, be written as

$$\begin{aligned} Y^E_{ij} &= \tau^E_{ij} + \eta^E_{ij} \quad , \; j = 1,\ldots,n^E_i \quad , \; i = 1,\ldots,k \quad , \\ Y^C_{ij} &= \tau^C_{ij} + \eta^C_{ij} \quad , \; j = 1,\ldots,n^C_i \quad , \; i = 1,\ldots,k \quad , \end{aligned} \tag{10}$$

where $\tau^E_{ij} = \beta_i\delta + \beta_i\gamma_i + \xi^E_{ij}$, and $\tau^C_{ij} = \beta_i\gamma_i + \xi^C_{ij}$. The reliability ρ_i of the response measure is defined as the pooled within-group ratio of the true score variance $\sigma^2_{\tau_i}$ to total variance $\sigma^2_{\tau_i} + \sigma^2_{\eta_i}$. Thus the reliability of the i^{th} response measure is $\rho_i = \sigma^2_{\tau_i}/(\sigma^2_{\tau_i} + \sigma^2_{\eta_i}) = \sigma^2_{\xi_i}/(\sigma^2_{\xi_i} + \sigma^2_{\eta_i})$.

Consider the population value of the standardized mean difference in two cases, one in which the measurements are error-free (i.e., $\sigma^2_\eta = 0$) and one in which errors of measurement are present (i.e., $\sigma^2_\eta \neq 0$). For simplicity of notation,

the subscript i denoting the particular experiment is omitted in the exposition that follows, but the results apply to each experiment when properly indexed. If there are no errors of measurement, then the within-cell standard deviation is simply σ_ξ^2. Let δ denote the population value of the standardized mean difference when there are no errors of measurement. Then

$$\delta = (\mu^E - \mu^C)/\sigma_\xi \quad , \tag{11}$$

where μ^E and μ^C are the population means of the experimental and control groups, respectively. Note that the use of the symbol δ in (11) is consistent with the definition of δ used in the structural model (1).

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

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

The relationship between δ and δ' can be expressed as

$$\delta' = \delta(\sigma_\xi \; / \; \sqrt{\sigma_\xi^2 + \sigma_\eta^2}) = \delta\sqrt{\rho} \quad , \tag{12}$$

where ρ is the reliability of the response measure.

5.3 Correcting Estimators for Measurement Error

From (12), the population value of the standardized mean difference depends explicitly on the reliability of the response measure. If the object is to estimate the value of δ, the standardized mean difference with no errors of measurement, then estimation of δ' instead of δ can result in biased estimates. Since reliabilities cannot exceed one, the effect of measurement error is to reduce the magnitude of the parameter δ' compared with δ. In particular, errors of measurement cause the estimator g^U to estimate δ' instead of δ, so that $E(g^U) = \delta' = \delta\sqrt{\rho}$. Hence errors of measurement result in underestimates of the parameter δ.

If the reliability ρ is known, the bias can be removed by dividing g^U by $\sqrt{\rho}$. When we combine several estimates that use response scales with different reliabilities, each esti-

mate can be corrected for measurement error separately. This leads to the estimator

$$G^{UR} = \frac{1}{k} \sum_{i=1}^{k} g_i^U / \sqrt{\rho_i} \quad , \tag{13}$$

where ρ_i is the reliability of the i^{th} response measure. The results of this section are summarized in the following theorem.

Theorem 4: Let g_i^U, $i = 1,\ldots,k$ be the unbiased estimators (7) arising from k independent experiments with an (error free) effect size δ. Then G^{UR} as defined by (13) is an unbiased estimator of δ with variance,

$$\mathrm{Var}(G^{UR}) = \frac{1}{k^2} \sum_{i=1}^{k} \left[\mathrm{Var}(g_i^U)/\rho_i \right] \quad .$$

6. THE EFFECT OF VALIDITY OF RESPONSE MEASURES

The effect of validity of response measures has not been considered in previous sections. The structural models (1) and (9) do not admit the possibility that some response measures have unique factors. For example, some experiments might use an expensive standardized test to measure reading achievement, whereas other studies use locally developed tests that are correlated with the standardized test. If the locally developed tests have unique factors, they will not be perfectly valid measures of reading achievement as measured by the standardized test. This section deals with the effect of invalidity of response measures on estimators of effect size. A structural model for invalidity is presented in Section 6.1. A simultaneous correction for the effects of invalidity and measurement error on estimator g^U is derived in Section 6.2, for the special case in which treatment does not affect unique factors. The effect of invalidity on the estimator g^U when treatment affects unique factors is derived in Section 6.3.

6.1 A Model for Invalidity

One model for test validity assumes that a collection of tests share a common factor, but that some tests also have unique factors. If the population of test scores on a particular test are generated by a model which includes both the common factor (among all the tests) and a unique factor, then the test is partially invalid. The (conceptual) structural

model for a collection of k response measures is therefore:

$$Y^E_{ij} = \beta_i\delta + \zeta_i + \beta_i\gamma_i + \xi^E_{ij} + \theta^E_{ij} + \eta^E_{ij},$$
$$j = 1,\ldots,n^E_i \quad , \quad i = 1,\ldots,k \quad , \tag{14}$$
$$Y^C_{ij} = \beta_i\gamma_i + \xi^C_{ij} + \theta^C_{ij} + \eta^C_{ij}, \; j = 1,\ldots,n^C_i, \; i = 1,\ldots,k,$$

where $\beta_i = \sigma_{\xi_i}$, $\beta_i\delta$ and ζ_i are the contributions of the treatment effect on common and the $\underline{i}^{th}$ unique factors, γ_i is a location (scale mean) parameter, ξ_{ij} and θ_{ij} are the contribution of subject-treatment interaction on the common and unique factors, and η_{ij} is an error of measurement. Note that all of the parameters of the model cannot be identified without additional information or restrictions. We usually want to estimate δ, the effect of treatment on the common factor, standardized by the common factor standard deviation within groups. Note that when $\zeta_i = \theta_{ij} = \eta_{ij} = 0$, $i = 1,\ldots,k$, $j = 1,\ldots,n_i$, model (14) reduces to model (1) and δ has the same interpretation as in model (1).

6.2 The Effect of Validity when Treatment Affects only the Common Factor

One of the restrictions that can be applied to model (14) is to assume that $\zeta_i = 0$ for each $i = 1,\ldots,k$, that is, that treatment affects Y only through the common factor. We also assume that ξ_{ij}, θ_{st}, and η_{uv} are mutually independent with zero expectations, which are standard assumptions from classical test theory. Under these assumptions, it is possible to obtain an unbiased estimate of the treatment effect δ if the correlation of each response measure with a valid response measure (i.e. a test with no unique factor) of known reliability is available. Let X_i, $i = 1,\ldots,k$, be a series of response measures with reliabilities ρ^R_i, $i = 1,\ldots,k$, and suppose that the X_i have no unique factors, but share the common factor of another (partially invalid) series of response measures Y_i, $i = 1,\ldots,k$. If the correlation, ρ^X_i, of X_i with Y_i is known, an unbiased estimate of δ in model (14) can be obtained from a series of measurements using the response scales Y_i, $i = 1,\ldots,k$. Under these assumptions, invalidity biases the estimators g^U_i given in (7) downward.

These results are summarized below.

Theorem 5: Let X_i, i = 1,...,k, be valid response scales with reliabilities ρ_i^R and let Y_i, i = 1,...,k, be response scales whose correlation with the X_i are given by ρ_i^X, as defined above. If g_i^U, i = 1,...,k are the estimators (7) based on the measures Y_i, i = 1,...,k, then estimators

$$g_i^{UV} = g_i^U \sqrt{\rho_i^R}/\rho_i^X \quad , \tag{15}$$

and

$$G^{UV} = \frac{1}{k} \sum_{i=1}^{k} g_i^{UV} \quad , \tag{16}$$

are unbiased estimators of δ in model (14), and the variance of G^{UV} is given by

$$\mathrm{Var}(G^{UV}) = \frac{1}{k^2} \sum_{i=1}^{k} \mathrm{Var}(g_i^U)\rho_i^R/(\rho_i^X)^2 \quad .$$

Proof: Write the standardized mean difference δ'' as a function of the parameters of model (14) and their moments, where for simplicity of notation, the subscript i referring to the i^{th} experiment is omitted. Recall that $\zeta = 0$, and the covariances of ξ_j, θ_s and η_t are zero, by hypothesis. Then

$$\delta'' = \frac{\beta\delta}{\sqrt{\beta^2 + \sigma_\theta^2 + \sigma_\eta^2}} \quad . \tag{17}$$

The validity coefficient ρ^V of the response measure can be expressed by its within-group correlation with the common factor. Because the common factor is a linear transformation of ξ, the validity coefficient can be written as $\rho^V = \rho_{\xi Y} = \beta/\sqrt{\beta^2 + \sigma_\xi^2 + \sigma_\eta^2}$. Therefore the population value of the standardized mean difference δ'' can be expressed as $\delta'' = \delta\rho_{\xi Y}$. It seems unlikely that the population correlation of an invalid measure with the common factor among a series of measures would be known. If X is a measure that is not perfectly reliable, shares the Y common factor but has no unique factor, then the correlation $\rho_{\xi Y}$ can be obtained from ρ_{XY} by the familiar disattenuation formula (see e.g., Lord & Novick, 1968):

$$\rho_{\xi Y} = \rho_{XY}/\sqrt{\rho}\ , \tag{18}$$

where ρ is the reliability of the measure X. Thus the population standardized mean difference δ'' can be written in terms of a correlation with a valid but unreliable measure X and the reliability ρ of X, namely,

$$\delta'' = \delta\rho_{XY}/\sqrt{\rho}\ . \tag{19}$$

Theorem 5 follows immediately from (19). ||

Since $\rho_{XY} \leq \sqrt{\rho}$ it follows that $\delta'' \leq \delta$. This means that invalidity always reduces the standardized mean difference when treatment affects only the common factor among the response measures.

6.3 The Effect of Invalidity When Treatment Affects Both Common and Unique Factors

When the treatment affects the response measure Y through both common and unique factors, the invalidity of the response measure may either increase or decrease the standardized mean difference. Hence no simple characterization of the effect of invalidity on estimates of δ obtained from the estimators g_i^U given by (7) is possible.

We omit the subscript i to denote a particular response measure, and assume that $\mathrm{Cov}(\xi,\theta) = \mathrm{Cov}(\xi,\eta) = \mathrm{Cov}(\theta,\eta) = 0$. Now ζ, the effect of treatment via the unique factor, is nonzero. The standardized mean difference under model (14) and these assumptions is denoted $\tilde{\delta}$. Specifically,

$$\tilde{\delta} = (\beta\delta + \zeta)/\sqrt{\beta^2 + \sigma_\theta^2 + \sigma_\eta^2}\ .$$

Note that if $\zeta = 0$, $\tilde{\delta}$ reduces to the standardized mean difference δ'' given by (17). If $\zeta = 0$ and $\sigma_\theta^2 = \sigma_\eta^2 = 0$, then $\tilde{\delta} = \delta$. If ζ, the treatment effect via the unique factor is large enough, namely if

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

then $\tilde{\delta} > \delta$. If $\zeta < \zeta_c$, $\tilde{\delta} < \delta$.

7. WEIGHTING THE ESTIMATES FROM SEVERAL EXPERIMENTS

In Sections 1 to 6, the properties of estimators that give equal weight to each experiment are considered. If the

experiments have very different sample sizes , the precision of the estimates obtained from the experiments will also be unequal because the variance of each estimator is a function of sample size and δ, the true effect size. In this case, the precision of the combined estimate of δ can be improved by giving more weight to experiments which contribute more precise information to the overall estimate. The optimal weights clearly depend on the variances of the estimators from each experiment, which in turn depend on the sample size for each experiment and on δ. Specifically, let G^{UW} be the weighted estimator

$$G^{UW} = \sum_{i=1}^{k} w_i g_i^U \quad , \tag{20}$$

where g_i^U, $i = 1,\ldots,k$ are the estimators (7), and $w_i > 0$.

The weights w_i, $i = 1,\ldots,k$ that minimize the variance of G^{UW} depend on δ in general, but an expression for these optimal weights is easy to obtain. When the experimental and control groups of each experiment are equal in size, that is, $n_i^E = n_i^C$, $i = 1,\ldots,k$, and when $n_i^E + n_i^C$ is large, the optimal weights have a simple approximation which is independent of δ. These results are summarized by the following fact.

Theorem 6: When g_i^U are the estimators (7) with variance v_i, $i = 1,\ldots,k$, and G^{UW} is the estimator $\sum_{i=1}^{k} w_i g_i^U$, then the weights that minimize $\mathrm{Var}[G^{UW}]$ are given by

$$w_i = \frac{1/v_i}{\sum_{j=1}^{k} (1/v_j)} \quad . \tag{21}$$

If $n_i^E = n_i^C$, $i = 1,\ldots,k$, and m_i is large for each $i = 1,\ldots,k$, then the weights that minimize the variance of G^{UW} are approximately

$$w_i = \frac{m_i}{\sum_{j=1}^{k} m_j} \quad , \tag{22}$$

where m_i is the number of degrees of freedom used to estimate

β_i for each g_i^U.

Proof: This fact is proved by minimizing $V(w_1,\ldots,w_k) = \mathrm{Var}(G^{UW}) = \sum_{i=1}^{k} w_i^2 v_i$, where $v_i = \mathrm{Var}(g_i^U)$ subject to the constraints that $w_i > 0$ for all i and $\sum_{i=1}^{k} w_i = 1$. The weights which minimize V are given by the expression (21). One approach to obtaining optimal weights is to use a sample estimate of δ in the formula for each v_i. A simpler approach when the experimental and control group sample sizes are equal and all the m_i are large is to use the large sample approximation to the variances v_i, namely $v_i = (2/m_i)(1 + \delta^2/4)$. In this case the second factor $(1 + \delta^2/4)$ cancels and the w_i are given approximately by $m_i / \sum_{j=1}^{k} m_j$. ||

8. EXAMPLE

The techniques described in this paper were applied to the results of an experiment in the field of research on teaching conducted by N. L. Gage. A sample of 33 third-grade classes in two school districts either received or did not receive a learning skills program designed to improve students' academic achievement. The two school districts administered different achievement tests which were to be used as the dependent variable. One district used the Science Research Associates (SRA) achievement test, while the other district used the Comprehensive Test of Basic Skills (CTBS) achievement test. The object of the analysis is to estimate the standardized treatment effect on the arithmetic achievement as measured by the arithmetic subscales of the tests.

The two tests have different means and variances so direct combination of test scores is not possible. Instead, we assume that the standardized treatment effect is the same in both districts, and use the methods of this paper to estimate effect size. The reliability of each class mean with the classroom as the unit of analysis is so high ($r > .99$) that the correction for unreliability is negligible. The analysis summarized in Table II shows that G^{UW}, the estimate of the treatment effect based on the weights given in (22) is .239. An approximate confidence interval for δ obtained by the

method described in Section 4 is

$$-.447 \leq \delta \leq .925.$$

TABLE II

Example of an Estimate of the Standardized Mean Difference.

	District I	District II
n^E	9	7
n^C	9	8
m	16	13
$\bar{Y}^E - \bar{Y}^C$	.154	.444
S	.979	1.201
$g = (\bar{Y}^E - \bar{Y}^C)/S$	.157	.370
$g^U = c(m)g$	.150	.348
$w = m/\Sigma m$	.552	.448
$\text{Var}\,(g^U) = (1/\tilde{n} + \hat{\delta}^2/2m)$	.224	.270

ACKNOWLEDGMENTS

This research was supported by a grant from the Spencer Foundation.

REFERENCES

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

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

Johnson, N.L., & Welch, B.L. Applications of the noncentral t-distribution. Biometrika, 1939, 31, 362-389.

Lehmann, E.L.,& Scheffé,H. Completeness, similar regions and unbiased estimation. Sankhyá, 1950, 10, 305-340.

Lord, F.L., & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Welsey, 1969.

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

Walker, G.A., & Saw, J.G. The distribution of linear combinations of t-variables. Journal of the American Statistical Association, 1978, 73, 876-878.

AUTHOR

HEDGES, LARRY V. Address: Department of Education, The University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637. Title: Assistant Professor. Degrees: B.A. University of California, San Diego; M.S., Ph.D. Stanford University. Specialization: Statistics.