CHAPTER 4

Effect Sizes Based on Means

INTRODUCTION

When the studies report means and standard deviations, the preferred effect size is usually the raw mean difference, the standardized mean difference, or the response ratio. These effect sizes are discussed in this chapter.

RAW (UNSTANDARDIZED) MEAN DIFFERENCE *D*

When the outcome is reported on a meaningful scale *and* all studies in the analysis use the same scale, the meta-analysis can be performed directly on the raw difference in means (henceforth, we will use the more common term, *raw mean difference*). The primary advantage of the raw mean difference is that it is intuitively meaningful, either inherently (for example, blood pressure, which is measured on a known scale) or because of widespread use (for example, a national achievement test for students, where all relevant parties are familiar with the scale).

Consider a study that reports means for two groups (Treated and Control) and suppose we wish to compare the means of these two groups. Let μ_1 and μ_2 be the true (population) means of the two groups. The population mean difference is defined as

$$\Delta = \mu_1 - \mu_2. \tag{4.1}$$

In the two sections that follow we show how to compute an estimate D of this parameter and its variance from studies that used two independent groups and from studies that used paired groups or matched designs.

Introduction to Meta-Analysis. Michael Borenstein, L. V. Hedges, J. P. T. Higgins and H. R. Rothstein
 ISBN: 978-0-470-05724-7

Computing *D* from studies that use independent groups

We can estimate the mean difference Δ from a study that used two independent groups as follows. Let $\overline{X}_1$ and $\overline{X}_2$ be the sample means of the two independent groups. The sample estimate of Δ is just the difference in sample means, namely

$$D = \overline{X}_1 - \overline{X}_2. \tag{4.2}$$

Note that uppercase D is used for the *raw* mean difference, whereas lowercase d will be used for the *standardized* mean difference (below).

Let S_1 and S_2 be the sample standard deviations of the two groups, and n_1 and n_2 be the sample sizes in the two groups. If we assume that the two population standard deviations are the same (as is assumed to be the case in most parametric data analysis techniques), so that $\sigma_1 = \sigma_2 = \sigma$, then the variance of D is

$$V_D = \frac{n_1 + n_2}{n_1 n_2} S^2_{pooled}, \tag{4.3}$$

where

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}. \tag{4.4}$$

If we don't assume that the two population standard deviations are the same, then the variance of D is

$$V_D = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}. \tag{4.5}$$

In either case, the standard error of D is then the square root of V,

$$SE_D = \sqrt{V_D}. \tag{4.6}$$

For example, suppose that a study has sample means $\overline{X}_1 = 103.00$, $\overline{X}_2 = 100.00$, sample standard deviations $S_1 = 5.5$, $S_2 = 4.5$, and sample sizes $n_1 = n_2 = 50$. The raw mean difference D is

$$D = 103.00 - 100.00 = 3.00.$$

If we assume that $\sigma_1 = \sigma_2$ then the pooled standard deviation within groups is

$$S_{pooled} = \sqrt{\frac{(50 - 1) \times 5.5^2 + (50 - 1) \times 4.5^2}{50 + 50 - 2}} = 5.0249.$$

The variance and standard error of D are given by

$$V_D = \frac{50 + 50}{50 \times 50} \times 5.0249^2 = 1.0100,$$

and

$$SE_D = \sqrt{1.0100} = 1.0050.$$

If we do not assume that $\sigma_1 = \sigma_2$ then the variance and standard error of D are given by

$$V_D = \frac{5.5^2}{50} + \frac{4.5^2}{50} = 1.0100$$

and

$$SE_D = \sqrt{1.0100} = 1.0050.$$

In this example formulas (4.3) and (4.5) yield the same result, but this will be true only if the sample size and/or the estimate of the variances is the same in the two groups.

Computing *D* from studies that use matched groups or pre-post scores

The previous formulas are appropriate for studies that use two independent groups. Another study design is the use of matched groups, where pairs of participants are matched in some way (for example, siblings, or patients at the same stage of disease), with the two members of each pair then being assigned to different groups. The unit of analysis is the pair, and the advantage of this design is that each pair serves as its own control, reducing the error term and increasing the statistical power. The magnitude of the impact depends on the correlation between (for example) siblings, with a higher correlation yielding a lower variance (and increased precision).

The sample estimate of Δ is just the sample mean difference, D. If we have the difference score for each pair, which gives us the mean difference $\overline{X}_{diff}$ and the standard deviation of these differences (S_{diff}), then

$$D = \overline{X}_{diff}, \tag{4.7}$$

$$V_D = \frac{S_{diff}^2}{n}, \tag{4.8}$$

where n is the number of pairs, and

$$SE_D = \sqrt{V_D}. \tag{4.9}$$

For example, if the mean difference is 5.00 with standard deviation of the difference of 10.00 and n of 50 pairs, then

$$D = 5.0000,$$

$$V_D = \frac{10.00^2}{50} = 2.0000, \tag{4.10}$$

and

$$SE_D = \sqrt{2.00} = 1.4142. \tag{4.11}$$

Alternatively, if we have the mean and standard deviation for each set of scores (for example, siblings *A* and *B*), the difference is

$$D = \overline{X}_1 - \overline{X}_2. \tag{4.12}$$

The variance is again given by

$$V_D = \frac{S_{diff}^2}{n}, \tag{4.13}$$

where n is the number of pairs, and the standard error is given by

$$SE_D = \sqrt{V_D}. \tag{4.14}$$

However, in this case we need to compute the standard deviation of the difference scores from the standard deviation of each sibling's scores. This is given by

$$S_{diff} = \sqrt{S_1^2 + S_2^2 - 2 \times r \times S_1 \times S_2} \tag{4.15}$$

where r is the correlation between 'siblings' in matched pairs. If $S_1 = S_2$, then (4.15) simplifies to

$$S_{diff} = \sqrt{2 \times S_{pooled}^2 (1 - r)}. \tag{4.16}$$

In either case, as r moves toward 1.0 the standard error of the paired difference will decrease, and when $r = 0$ the standard error of the difference is the same as it would be for a study with two independent groups, each of size n.

For example, suppose the means for siblings A and B are 105.00 and 100.00, with standard deviations 10 and 10, the correlation between the two sets of scores is 0.50, and the number of pairs is 50. Then

$$D = 105.00 - 100.00 = 5.0000,$$

$$V_D = \frac{10.00^2}{50} = 2.0000,$$

and

$$SE_D = \sqrt{2.00} = 1.4142.$$

In the calculation of V_D, the S_{diff} is computed using

$$S_{diff} = \sqrt{10^2 + 10^2 - 2 \times 0.50 \times 10 \times 10} = 10.0000$$

or

$$S_{diff} = \sqrt{2 \times 10^2 (1 - 0.50)} = 10.0000.$$

The formulas for matched designs apply to pre-post designs as well. The pre and post means correspond to the means in the matched groups, n is the number of subjects, and r is the correlation between pre-scores and post-scores.

Calculation of effect size estimates from information that is reported

When a researcher has access to a full set of summary data such as the mean, standard deviation, and sample size for each group, the computation of the effect size and its variance is relatively straightforward. In practice, however, the researcher will often be working with only partial data. For example, a paper may publish only the *p*-value, means and sample sizes from a test of significance, leaving it to the meta-analyst to back-compute the effect size and variance. For information on computing effect sizes from partial information, see Borenstein *et al*. (2009).

Including different study designs in the same analysis

Sometimes a systematic review will include studies that used independent groups and also studies that used matched groups. From a statistical perspective the effect size (D) has the same meaning regardless of the study design. Therefore, we can compute the effect size and variance from each study using the appropriate formula, and then include all studies in the same analysis. While there is no technical barrier to using different study designs in the same analysis, there may be a concern that studies which used different designs might differ in substantive ways as well (see Chapter 40).

For all study designs (whether using independent or paired groups) the direction of the effect ($\overline{X}_1 - \overline{X}_2$ or $\overline{X}_2 - \overline{X}_1$) is arbitrary, except that the researcher must decide on a convention and then apply this consistently. For example, if a positive difference will indicate that the treated group did better than the control group, then this convention must apply for studies that used independent designs and for studies that used pre-post designs. In some cases it might be necessary to reverse the computed sign of the effect size to ensure that the convention is followed.

STANDARDIZED MEAN DIFFERENCE, *d* AND *g*

As noted, the raw mean difference is a useful index when the measure is meaningful, either inherently or because of widespread use. By contrast, when the measure is less well known (for example, a proprietary scale with limited distribution), the use of a raw mean difference has less to recommend it. In any event, the raw mean difference is an option only if all the studies in the meta-analysis use the same scale. If different studies use different instruments (such as different psychological or educational tests) to assess the outcome, then the scale of measurement will differ from study to study and it would not be meaningful to combine raw mean differences.

In such cases we can divide the mean difference in each study by that study's standard deviation to create an index (the standardized mean difference) that would be comparable across studies. This is the same approach suggested by Cohen (1969, 1987) in connection with describing the magnitude of effects in statistical power analysis.

The standardized mean difference can be considered as being comparable across studies based on either of two arguments (Hedges and Olkin, 1985). If the outcome measures in all studies are linear transformations of each other, the standardized mean difference can be seen as the mean difference that would have been obtained if all data were transformed to a scale where the standard deviation within-groups was equal to 1.0.

The other argument for comparability of standardized mean differences is the fact that the standardized mean difference is a measure of overlap between distributions. In this telling, the standardized mean difference reflects the difference between the distributions in the two groups (and how each represents a distinct cluster of scores) even if they do not measure exactly the same outcome (see Cohen, 1987, Grissom and Kim, 2005).

Consider a study that uses two independent groups, and suppose we wish to compare the means of these two groups. Let μ_1 and σ_1 be the true (population) mean and standard deviation of the first group and let μ_2 and σ_2 be the true (population) mean and standard deviation of the other group. If the two population standard deviations are the same (as is assumed in most parametric data analysis techniques), so that $\sigma_1 = \sigma_2 = \sigma$, then the standardized mean difference parameter or population standardized mean difference is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}. \tag{4.17}$$

In the sections that follow, we show how to estimate δ from studies that used independent groups, and from studies that used pre-post or matched group designs. It is also possible to estimate δ from studies that used other designs (including clustered designs) but these are not addressed here (see resources at the end of this Part). We make the common assumption that ${\sigma_1}^2 = {\sigma_2}^2$, which allows us to pool the estimates of the standard deviation, and do not address the case where these are assumed to differ from each other.

Computing *d* and *g* from studies that use independent groups

We can estimate the standardized mean difference (δ) from studies that used two independent groups as

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{within}}. \tag{4.18}$$

In the numerator, $\overline{X}_1$ and $\overline{X}_2$ are the sample means in the two groups. In the denominator S_{within} is the within-groups standard deviation, pooled across groups,

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \tag{4.19}$$

where n_1 and n_2 are the sample sizes in the two groups, and S_1 and S_2 are the standard deviations in the two groups. The reason that we pool the two sample

estimates of the standard deviation is that even if we assume that the underlying population standard deviations are the same (that is $\sigma_1 = \sigma_2 = \sigma$), it is unlikely that the sample estimates S_1 and S_2 will be identical. By pooling the two estimates of the standard deviation, we obtain a more accurate estimate of their common value.

The sample estimate of the standardized mean difference is often called Cohen's *d* in research synthesis. Some confusion about the terminology has resulted from the fact that the index δ, originally proposed by Cohen as a *population parameter* for describing the size of effects for statistical power analysis is also sometimes called *d*. In this volume we use the symbol δ to denote the effect size parameter and *d* for the sample estimate of that parameter.

The variance of *d* is given (to a very good approximation) by

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}. \tag{4.20}$$

In this equation the first term on the right of the equals sign reflects uncertainty in the estimate of the mean difference (the numerator in (4.18)), and the second reflects uncertainty in the estimate of S_{within} (the denominator in (4.18)).

The standard error of *d* is the square root of V_d,

$$SE_d = \sqrt{V_d}. \tag{4.21}$$

It turns out that *d* has a slight bias, tending to overestimate the absolute value of δ in small samples. This bias can be removed by a simple correction that yields an unbiased estimate of δ, with the unbiased estimate sometimes called Hedges' *g* (Hedges, 1981). To convert from *d* to Hedges' *g* we use a correction factor, which is called *J*. Hedges (1981) gives the exact formula for *J*, but in common practice researchers use an approximation,

$$J = 1 - \frac{3}{4df - 1}. \tag{4.22}$$

In this expression, *df* is the degrees of freedom used to estimate S_{within}, which for two independent groups is $n_1 + n_2 - 2$. This approximation always has error of less than 0.007 and less than 0.035 percent when $df \geq 10$ (Hedges, 1981). Then,

$$g = J \times d, \tag{4.23}$$

$$V_g = J^2 \times V_d, \tag{4.24}$$

and

$$SE_g = \sqrt{V_g}. \tag{4.25}$$

For example, suppose a study has sample means $\overline{X}_1 = 103$, $\overline{X}_2 = 100$, sample standard deviations $S_1 = 5.5$, $S_2 = 4.5$, and sample sizes $n_1 = n_2 = 50$. We would estimate the pooled-within-groups standard deviation as

$$S_{within} = \sqrt{\frac{(50-1)\times 5.5^2 + (50-1)\times 4.5^2}{50+50-2}} = 5.0249.$$

Then,

$$d = \frac{103-100}{5.0249} = 0.5970,$$

$$V_d = \frac{50+50}{50\times 50} + \frac{0.5970^2}{2(50+50)} = 0.0418,$$

and

$$SE_d = \sqrt{0.0418} = 0.2044.$$

The correction factor (J), Hedges' g, its variance and standard error are given by

$$J = \left(1 - \frac{3}{4\times 98 - 1}\right) = 0.9923,$$

$$g = 0.9923 \times 0.5970 = 0.5924,$$

$$v_g = 0.9923^2 \times 0.0418 = 0.0411,$$

and

$$SE_g = \sqrt{0.0411} = 0.2028.$$

The correction factor (J) is always less than 1.0, and so g will always be less than d in absolute value, and the variance of g will always be less than the variance of d. However, J will be very close to 1.0 unless df is very small (say, less than 10) and so (as in this example) the difference is usually trivial (Hedges, 1981).

Some slightly different expressions for the variance of d (and g) have been given by different authors and even the same authors at different times. For example, the denominator of the second term of the variance of d is given here as $2(n_1 + n_2)$. This expression is obtained by one method (assuming the n's become large with δ fixed). An alternate derivation (assuming n's become large with $\sqrt{n}\delta$ fixed) leads to a denominator in the second term that is slightly different, namely $2(n_1 + n_2 - 2)$. Unless n_1 and n_2 are very small, these expressions will be almost identical.

Similarly, the expression given here for the variance of g is J^2 times the variance of d, but many authors ignore the J^2 term because it is so close to unity in most cases. Again, while it is preferable to include this correction factor, the inclusion of this factor is likely to make little practical difference.

Computing *d* and *g* from studies that use pre-post scores or matched groups

We can estimate the standardized mean difference (δ) from studies that used matched groups or pre-post scores in one group. The formula for the sample estimate of d is

$$d = \frac{\overline{Y}_{diff}}{S_{within}} = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{within}}. \tag{4.26}$$

This is the same formula as for independent groups (4.18). However, when we are working with independent groups the natural unit of deviation is the standard deviation within groups and so this value is typically reported (or easily imputed). By contrast, when we are working with matched groups, the natural unit of deviation is the standard deviation *of the difference scores*, and so *this* is the value that is likely to be reported. To compute d from the standard deviation of the differences we need to impute the standard deviation within groups, which would then serve as the denominator in (4.26).

Concretely, when working with a matched study, the standard deviation within groups can be imputed from the standard deviation of the difference, using

$$S_{within} = \frac{S_{diff}}{\sqrt{2(1-r)}}, \tag{4.27}$$

where r is the correlation between pairs of observations (e.g., the pretest-posttest correlation). Then we can apply (4.26) to compute d. The variance of d is given by

$$V_d = \left(\frac{1}{n} + \frac{d^2}{2n}\right) 2(1-r), \tag{4.28}$$

where n is the number of pairs. The standard error of d is just the square root of V_d,

$$SE_d = \sqrt{V_d}\,. \tag{4.29}$$

Since the correlation between pre- and post-scores is required to impute the standard deviation within groups from the standard deviation of the difference, we must assume that this correlation is known or can be estimated with high precision. Otherwise we may estimate the correlation from related studies, and possibly perform a sensitivity analysis using a range of plausible correlations.

To compute Hedges' g and associated statistics we would use formulas (4.22) through (4.25). The degrees of freedom for computing J is $n - 1$, where n is the number of pairs.

For example, suppose that a study has pre-test and post-test sample means $\overline{X}_1 = 103$, $\overline{X}_2 = 100$, sample standard deviation of the difference $S_{diff} = 5.5$, sample size $n = 50$, and a correlation between pre-test and post-test of $r = 0.7$. The standard deviation within groups is imputed from the standard deviation of the difference by

$$S_{within} = \frac{5.5}{\sqrt{2(1-0.7)}} = 7.1005.$$

Then d, its variance and standard error are computed as

$$d = \frac{103 - 100}{7.1000} = 0.4225,$$

$$v_d = \left(\frac{1}{50} + \frac{0.4225^2}{2 \times 50}\right)(2(1 - 0.7)) = 0.0131,$$

and

$$SE_d = \sqrt{0.0131} = 0.1143.$$

The correction factor J, Hedges' g, its variance and standard error are given by

$$J = \left(1 - \frac{3}{4 \times 49 - 1}\right) = 0.9846,$$

$$g = 0.9846 \times 0.4225 = 0.4160,$$

$$V_g = 0.9846^2 \times 0.0131 = 0.0127,$$

and

$$SE_g = \sqrt{0.0127} = 0.1126.$$

Including different study designs in the same analysis

As we noted earlier, a single systematic review can include studies that used independent groups and also studies that used matched groups. From a statistical perspective the effect size (d or g) has the same meaning regardless of the study design. Therefore, we can compute the effect size and variance from each study using the appropriate formula, and then include all studies in the same analysis. While there are no technical barriers to using studies with different designs in the same analysis, there may be a concern that these studies could differ in substantive ways as well (see Chapter 40).

For all study designs the direction of the effect ($\overline{X}_1 - \overline{X}_2$ or $\overline{X}_2 - \overline{X}_1$) is arbitrary, except that the researcher must decide on a convention and then apply this consistently. For example, if a positive difference indicates that the treated group did better than the control group, then this convention must apply for studies that used independent designs and for studies that used pre-post designs. It must also apply for all outcome measures. In some cases (for example, if some studies defined outcome as the number of correct answers while others defined outcome as the number of mistakes) it will be necessary to reverse the computed sign of the effect size to ensure that the convention is applied consistently.

RESPONSE RATIOS

In research domains where the outcome is measured on a physical scale (such as length, area, or mass) and is unlikely to be zero, the ratio of the means in the two groups might serve as the effect size index. In experimental ecology this effect size index is called the response ratio (Hedges, Gurevitch, & Curtis, 1999). It is important to recognize that the response ratio is only meaningful when the outcome

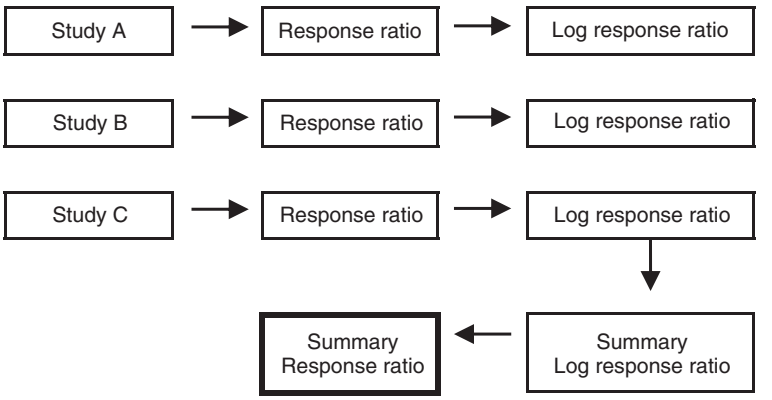

Figure 4.1 Response ratios are analyzed in log units.

is measured on a true ratio scale. The response ratio is not meaningful for studies (such as most social science studies) that measure outcomes such as test scores, attitude measures, or judgments, since these have no natural scale units and no natural zero points.

For response ratios, computations are carried out on a log scale (see the discussion under risk ratios, below, for an explanation). We compute the log response ratio and the standard error of the log response ratio, and use these numbers to perform all steps in the meta-analysis. Only then do we convert the results back into the original metric. This is shown schematically in Figure 4.1.

The response ratio is computed as

$$R = \frac{\overline{X}_1}{\overline{X}_2} \tag{4.30}$$

where $\overline{X}_1$ is the mean of group 1 and $\overline{X}_2$ is the mean of group 2. The log response ratio is computed as

$$lnR = \ln(R) = \ln\left(\frac{\overline{X}_1}{\overline{X}_2}\right) = \ln\left(\overline{X}_1\right) - \ln\left(\overline{X}_2\right). \tag{4.31}$$

The variance of the log response ratio is approximately

$$V_{lnR} = S^2_{pooled}\left(\frac{1}{n_1\left(\overline{X}_1\right)^2} + \frac{1}{n_2\left(\overline{X}_2\right)^2}\right), \tag{4.32}$$

where S_{pooled} is the pooled standard deviation. The approximate standard error is

$$SE_{\ln R} = \sqrt{V_{lnR}}. \tag{4.33}$$

Note that we do not compute a variance for the response ratio in its original metric. Rather, we use the *log* response ratio and its variance in the analysis to yield

a summary effect, confidence limits, and so on, in log units. We then convert each of these values back to response ratios using

$$R = \exp(lnR), \tag{4.34}$$

$$LL_R = \exp(LL_{lnR}), \tag{4.35}$$

and

$$UL_R = \exp(UL_{lnR}), \tag{4.36}$$

where *LL* and *UL* represent the lower and upper limits, respectively.

For example, suppose that a study has two independent groups with means $\overline{X}_1 = 61.515$, $\overline{X}_2 = 51.015$, pooled within-group standard deviation 19.475, and sample size $n_1 = n_2 = 10$.

Then *R*, its variance and standard error are computed as

$$R = \frac{61.515}{51.015} = 1.2058,$$

$$lnR = \ln(1.2058) = 0.1871,$$

$$V_{lnR} = 19.475^2 \left(\frac{1}{10 \times (61.515)^2} + \frac{1}{10 \times (51.015)^2} \right) = 0.0246.$$

and

$$SE_{lnR} = \sqrt{0.0246} = 0.1581.$$

SUMMARY POINTS

- The raw mean difference (*D*) may be used as the effect size when the outcome scale is either inherently meaningful or well known due to widespread use. This effect size can only be used when all studies in the analysis used precisely the same scale.
- The standardized mean difference (*d* or *g*) transforms all effect sizes to a common metric, and thus enables us to include different outcome measures in the same synthesis. This effect size is often used in primary research as well as meta-analysis, and therefore will be intuitive to many researchers.
- The response ratio (*R*) is often used in ecology. This effect size is only meaningful when the outcome has a natural zero point, but when this condition holds, it provides a unique perspective on the effect size.
- It is possible to compute an effect size and variance from studies that used two independent groups, from studies that used matched groups (or pre-post designs) and from studies that used clustered groups. These effect sizes may then be included in the same meta-analysis.