# 7
# Effect Sizes and Confidence Intervals[1]

**Geoff Cumming and Fiona Fidler**

An *effect size* (ES) is simply an amount of something of interest. It can be as simple as a mean, a percentage increase, or a correlation; or it may be a standardized measure of a difference, a regression weight, or the percentage of variance accounted for. Most research questions in the social sciences are best answered by finding estimated ESs, meaning *point estimates* of the true ESs in the population. Grissom and Kim (2005) provided a comprehensive discussion of ESs, and ways to calculate ES estimates.

A *confidence interval* (CI), most commonly a 95% CI, is an *interval estimate* of a population ES. It is an interval that extends above and below the point ES estimate; it indicates the *precision* of the point estimate. The *margin of error* (MOE) is the length of one arm of a CI. The most common CIs are symmetric, and for these the MOE is half the total width of the CI. The MOE is our measure of precision. Cumming (2007a) provided a brief overview of CIs and their meaning, and Cumming and Finch (2005)[2] provided an introduction, explained the advantages of CIs, and described a number of *rules of eye* to assist the understanding and interpretation of CIs. Smithson (2002) provided a more detailed account of CIs.

In the social sciences, statistical analysis is still dominated by null hypothesis significance testing (NHST). However, there is extensive evidence that NHST is poorly understood, frequently misused, and often leads to incorrect conclusions. It is an urgent research priority that social scientists shift from relying mainly on NHST to using better techniques, particularly those including ESs, CIs, and meta-analysis (MA). The best reference on statistics reform is the excellent book by Kline (2004). Wilkinson and the Task Force on Statistical Inference (TFSI) (1999) is a further source of good advice. Our aim in this chapter is to assist authors and manuscript reviewers to make the vital transition from over-reliance on NHST to more informative methods, including ESs, CIs, and MA.

## 1. Formulation of Main Questions as Estimation

An astronomer wishes to know the age of the Earth; a chemist measures the boiling point of an interesting new substance. These are the typical questions of science. Correspondingly, in the social sciences we wish to estimate how seriously divorce disrupts adolescent development, or the effect of a type of psychotherapy on depression in the elderly. The chemist reports her result as, for example, 27.35 ± 0.02°C, which signals that 27.35 is the point estimate of the boiling point, and 0.02 is the

Table 7.1 Desiderata for Effect Sizes and Confidence Intervals

| *Desideratum* | *Manuscrip Section(s)** |
|---|---|
| 1. The main questions to be addressed are formulated in terms of estimation and not simply null hypothesis significance testing. | I |
| 2. Previous research literature is discussed in terms of effect sizes, confidence intervals, and from a meta-analytic perspective. | I |
| 3. The rationale for the experimental design and procedure is explained and justified in terms of appropriateness for obtaining precise estimates of the target effect sizes. | I, M |
| 4. The dependent variables are described and operationalized with the aim that they should lead to good estimates of the target effect sizes. | M |
| 5. Results are presented and analyzed in terms of point estimates of the effect sizes. | R |
| 6. The precision of effect sizes is presented and analyzed in terms of confidence intervals. | R |
| 7. Wherever possible, results are presented in figures, with confidence intervals. | R, D |
| 8. Effect sizes are given substantive interpretation. | D |
| 9. Confidence intervals are given substantive interpretation. | D |
| 10. Meta-analytic thinking is used to interpret and discuss the findings. | D |

* *Note*: I = Introduction, M = Methods, R = Results, D = Discussion

precision of that estimate. Correspondingly, it is most informative if the psychologist reports the effect of the psychotherapy as an ES—the best estimate of the amount of change the therapy brings about—and a CI (e.g., 95%) to indicate the precision of that estimate. This approach can be contrasted with the less informative dichotomous thinking (there is, or is not, an effect) that results from NHST.

In expressing their aims, authors should use language such as:

- We estimate the extent of ...
- We will assess the influence of ... on ...
- Our aim is to find how large an effect ... has on ...
- We investigate the nature of the relation between ... and ...
- We will estimate how well our model fits these data ...

Expressions like these naturally lead to answers that are ES estimates. Contrast these with statements like, "We investigate whether the new treatment is better or worse than the old"; "We examine whether there is a relation between ... and ...." These statements suggest that a mere dichotomous yes-or-no answer would suffice. Almost certainly the new treatment has *some* effect; our real concern is whether that effect is tiny, or even negative, or is positive and usefully large. It is an estimate of ES that answers these latter questions.

Examine the wording used to express the aims and main questions of the manuscript, especially in the abstract and introduction, but also in the title. Replace any words that convey dichotomous thinking with words that ask for a quantitative answer.

## 2. Previous Literature

Traditionally, reviews of past research in the social sciences have focused on whether previously published studies have, or have not, found an effect. A review, or the Introduction section to a manuscript, may reduce to a mere list of studies that found a statistically significant effect, versus those whose results failed to reach statistical significance. That is an impoverished and even misleading approach, which ignores the sizes of effects observed, and the fact that many negative results are likely to have been Type II errors attributable to low statistical power.

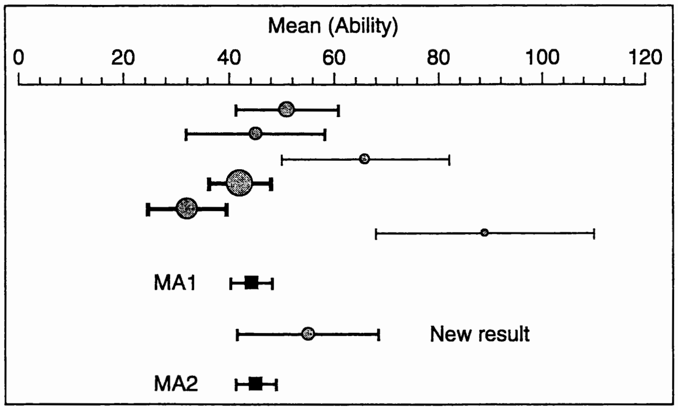
The American Psychological Association (APA) *Publication Manual* stated, "It is almost always necessary to include some measure of effect size in the Results section" (APA, 2010, p. 34). If an educational intervention increased mean reading age by 6.5 months, then 6.5 is our point estimate of the ES for that intervention. The *Manual* further stated, "Confidence intervals ... can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision ..., they are, in general, the best reporting strategy" (APA, p. 34). The above educational study may have found the increase to be 6.5 months of reading age, with a 95% CI of [6.5 ± 3.0], or [3.5, 9.5]. Narrow CIs indicate more precise estimates, and so the shorter a CI the better.

Past research should, wherever possible, be discussed in terms of the point and interval estimates obtained for the effects of interest. To researchers in many disciplines, that would not need stating. In the social sciences, however, many articles rely heavily on NHST, omit vital information about ES estimates, and conclude only that some intervention did or did not make a statistically significant difference (e.g., $p < .05$). Penetrating criticisms of NHST have been published by leading social scientists over more than half a century (e.g., Meehl, 1978), and advocacy of alternative techniques has been growing in volume and detail, especially in recent years (e.g., Fidler & Cumming, 2007). Other disciplines, notably medicine, have traveled at least part way along the road of statistical reform (Fidler, Cumming, Burgman, & Thomason, 2004). Kline (2004) identified 13 erroneous beliefs about *p* values and their use, and explained why NHST causes so much damage. He summarized the statistical reform debate, and proposed a well-informed and balanced approach to improving statistical practice by shifting from reliance on NHST to widespread use of ESs and CIs, together with MA. We recommend Kline's book; its position is similar to the position we take in this chapter.

Point estimates of ESs encompass a diversity of types of measures. Most simply, an ES is a mean or other measurement in the original measurement units: The average extent of masked priming was 27 ms; the mean improvement after therapy was 8.5 points on the Beck Depression Inventory; the regression of annual income against time spent in education was 3,700 dollars/year. Alternatively, an ES measure may be unit-free: After therapy, 48% of patients no longer met the criteria for the initial clinical diagnosis; the correlation between hours of study and final grade was .52; the odds ratio for risk of unemployment in young adults not in college is 1.4, for males compared with females. Some ES measures indicate percentage of variance accounted for, such as $R^2$, as often reported in multiple regression (Howell, 2002, Ch. 15), and $\eta^2$ or $\omega^2$, as often reported with analysis of variance (Howell, Ch. 11). An important class of ES measures are *standardized* ESs, including Cohen's *d* and Hedges' *g*, which are differences—typically between an experimental and a control group—measured in units of some relevant standard deviation (SD), for example the pooled SD of the two groups. Cumming and Finch (2001) explained Cohen's *d* and how to calculate CIs for *d*. Grissom and Kim (2005) is an excellent source of assistance with the calculation and presentation of a wide variety of ES measures.

The introduction to the manuscript should focus on the ES estimates reported in past research, to provide a setting for the results to be reported. It is often helpful to combine the past estimates, and meta-analysis (MA) allows that to be done quantitatively. Hunt (1997) gave a general introduction to MA, and explanation of its importance. Lipsey and Wilson (2001) provided an overview of how a MA should be conducted, and Chapter 19 of this volume discusses MA in more detail.

Figure 7.1 is a *forest plot*, which presents the hypothetical results of six past studies, and their combination by MA. The result of each study is shown as a point estimate, with its CI. The result of the MA is a weighted combination of the separate point estimates, also shown with its CI. This CI on the result is usually much shorter, indicating greater precision, as we would expect given that results are being combined over multiple studies. Some medical journals now routinely require the introduction to



*Note:* A forest plot, showing the results of six fictitious studies, each as a mean ability score, with 95% CI. These are the upper six dots, whose sizes and line thicknesses indicate sample size and study variance: Large dots and heavy lines signal large sample size and small variance and, therefore, a high weighting in the meta-analysis. MA1 is the weighted combination mean for those six studies, with its 95% CI. The hypothetical new result is similarly displayed as a mean and CI, with size again indicating weight in the meta-analysis. MA2 is the weighted combination mean for all seven studies, with its 95% CI.

**Figure 7.1** A Forest Plot.

each article to cite a MA—or wherever possible to carry out and report a new MA if none is available—as part of the justification for undertaking new research. That is a commendable requirement. Forest plots summarize a body of research in a compact and clear way; they are becoming common and familiar in medicine, and should be used more widely. Considering a number of estimates in combination, eventually including new results, is *meta-analytic thinking*, which we discuss below.

## 3. Experimental Design and the Precision of Estimates

Traditionally, statistical power estimates have been used to guide selection of the sample size *N* required if a planned study is to have a reasonable chance of identifying effects of a specified size (assuming they exist). This approach requires an estimate of error variability, preferably based on past research or a pilot study, and specification of what size of effect is likely, or is of theoretical interest. The TFSI stated that, "Because power computations are most meaningful when done before data are collected ..., it is important to show how effect-size estimates [to be used in power calculations] have been derived from previous research and theory" (Wilkinson et al., 1999, p. 596).

The power approach was advocated by Jacob Cohen, and his book (Cohen, 1988) provided tables and advice (see also Chapter 24, this volume). An Internet search readily identifies freely available software to carry out power calculations, including G*Power (http://www.psycho.uni-duesseldorf.de/aap/ projects/gpower/). The power approach can be useful, but statistical power is defined in the context of NHST, and has meaning only in relation to a specified null hypothesis. Null hypotheses are almost always statements of zero effect, zero difference, or zero change. Rarely is such a null hypothesis at all plausible, and so it a great advantage of CIs that no null hypothesis need be formulated. In addition, CIs offer an improved way to do the job of statistical power.

The TFSI recognized that CI width, or the MOE, is the appropriate measure of precision, or of the sensitivity of the experiment: "Once the study is analyzed, confidence intervals replace calculated power in describing results" (Wilkinson et al., 1999, p. 596). An important advance in statistical practice is routine use of precision, meaning the MOE, in planning a study, as well as in discussion and

ı mean
nd line
ʼy lines
ɡ in the
ɜs, with
ɜnd CI,
ɜighted

MA if none is available—
lable requirement. Forest
becoming common and
›er of estimates in combi-
e discuss below.

ion of the sample size N
; effects of a specified size
y, preferably based on past
ır is of theoretical interest.
when done before data are
ı power calculations] have
›. 596).
hen, 1988) provided tables
y identifies freely available
w.psycho.uni-duesseldorf.
ical power is defined in the
›thesis. Null hypotheses are
ırely is such a null hypothe-
ɜsis need be formulated. In

asure of precision, or of the
intervals replace calculated
ırtant advance in statistical
, as well as in discussion and

interpretation of results. Di Stefano, Fidler, and Cumming (2005) described such an alternative approach that avoids NHST and the need to choose a null hypothesis. It is based on calculation of what sample size is needed to give a CI a chosen expected width: How large must *N* be for the expected 95% CI to be no wider than, for example, 60 ms? Given a chosen experimental design, what sample size is needed for the expected MOE to be 0.2 units of Cohen's *d*? As with power, an estimate of variability is usually required, but no null hypothesis need be stated.

Justification of the experimental design and chosen sample size should appear as part of the rationale at the end of the Introduction section, or in the Methods section. It is often omitted from journal articles, having been overlooked by authors and reviewers, or squeezed out by strict word limits. Providing such justification is, however, especially important in cases where using too small a sample is likely to give estimates so imprecise that the research is scarcely worth doing, and may give misleading results. It is ethically problematic to carry out studies likely to give such inaccurate results. The converse—studies with such a large sample of participants that effects are estimated with greater precision than is necessary—are also ethically problematic, although these tend to be less common. The best way to justify a proposed design and sample size is in terms of the precision of estimates—the expected MOE—likely to be given by the results.

## 4. Dependent Variables

Specifying the experimental questions in terms of estimation of ESs leads naturally to choice of the dependent variables (DVs), or measures, that are most appropriate for those questions. Choose the operationalization of each DV that is most appropriate for expressing the ESs to be estimated, and that has adequate measurement properties, including reliability and validity. The aim is to choose measures that (1) relate substantively most closely to the experimental questions, and therefore will give results that are meaningful and interpretable, and (2) are most likely to give precise estimates of the targeted population effects.

In the Introduction section there may be discussion of methods used in past research, and this may help guide the choice of measures. In the Methods section there may be reference to published articles that provide information about the development of particular measures, and their psychometric properties. One important consideration is that the results to be reported should be as comparable as possible with previous research, and likely future research, so that meta-analytic combination over studies is as easy as possible. It can of course be a notable contribution to develop and validate an improved measure, but other things being equal it is advantageous to use measures already established in a field of research.

Choice of measures is partly a technical issue, with guidance provided by psychometric evidence of reliability and validity in the context of the planned experiment. It is also, and most importantly, a substantive issue that requires expert judgment by the researchers: The measures must tap the target concepts, and must give estimates of effects that can be given substantive and useful interpretation in the research context.

## 5. Results: Effect Sizes

The main role of the Results section is to report the estimated ESs that are the primary outcomes of the research. We mentioned in Desideratum 2 the wide range of possible ES measures, and emphasized that many of these are as simple and familiar as means, percentages, and correlations. In many cases it is possible to transform one ES measure into a number of others; Kirk (1996, 2003) provided formulas for this purpose. A correlation, for example, can be transformed into a value of Cohen's *d*. It is a routine part of MA to have to transform ES estimates reported in a variety of ways into some common measure, as the basis for conducting the MA. In medicine, odds ratio or log odds ratio are frequently used as the common ES measure, but in the social sciences Cohen's *d*, or some other standardized measure of difference (such as Hedges' *g*) is more frequently chosen as the basis for MA.

The authors of a manuscript need to consider which ES measures to report, bearing in mind ease of substantive interpretation, and the needs of future researchers wishing to include the results in some future MA. Often it will be best to present results in the original measurement scale of a DV, for simplicity and ease of interpretation, and also in some standardized form to assist both the comparison of results over different studies and the conduct of future MA. For example, an improvement in depression scores might be reported as mean change in score on the Beck Depression Inventory (BDI) because such scores are well known and easily interpreted by researchers and practitioners in the field. However, if the improvement is also reported as a standardized score the result is easily compared with, or combined with, the results of other studies of therapy, even where they have used other measures of depression. Similarly, a regression coefficient could be reported both in raw form, to assist understanding and interpretation, and as a standardized value, to assist comparison across different measures and different studies. In any case, it is vital to report SDs, and mean square error values, so that later meta-analysts have sufficient information to calculate whichever standardized ES measures they require.

A standardized measure of difference, such as Cohen's *d*, can be considered simply as a number of standard deviations. It is in effect a *z* score. It is important to consider which SD is most appropriate to use as the basis for standardization. Scores on the BDI, and changes in BDI scores, could be standardized against a published SD for the BDI. The SD unit would then be the extent of variation in some BDI reference population. That SD would have the advantage of being a stable and widely available value. Similarly, many IQ measures are already standardized to have a SD of 15. Alternatively, a change in BDI score could be expressed in units of the pre-test SD in a sample of participants. That would be a unit idiosyncratic to a specific study, and containing sampling error, but it might be chosen because it applies to the particular patient population being studied, rather than the BDI reference population. As so often is the case in research, informed judgment is needed to guide the choice of SD for standardization. When a manuscript reports a Cohen's *d* value, or any other standardized measure, it is essential that it makes clear what basis was chosen for standardization; when any reader interprets a standardized measure it is critical to have clearly in mind what SD units are being used.

It may be objected that much research has the aim not of estimating how large an effect some intervention has, but of testing a theory. However, theory testing is most informative if considered as a question of estimating goodness of fit, rather than of rejecting or not rejecting a hypothesis derived from the theory. A goodness of fit index, which may be a percentage of variance, or some other measure of distance between theoretical predictions and data, is an ES measure, and point and interval estimates of goodness of fit provide the best basis for evaluating how well the theory accounts for the data (Velicer et al., 2008).

## 6. Results: Confidence Intervals

Wilkinson et al. (1999) advised that: "Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients of association or variation whenever possible" (p. 599). The *Publication Manual* (APA, 2010) also recommended CIs: "Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size" (p. 34) It specified (p. 117) the following style for reporting CIs in text.

At the first occurrence in a paragraph write: "The mean decrease was 34.5 m, 95% CI: [12.0, 57.0], and so. ..." On later occasions in the paragraph, if the meaning is clear write simply: "The mean was 4.7 cm [–0.8, 10.2], which implies that ...," or, "The means were 84% [73, 92] and 65% [53, 76], respectively ...," or, "The correlation was .41 [.16, .61]. ..." The units should not be repeated inside the square brackets. Note that in the last example, which gives the 95% CI on Pearson's $r = .41$, for $N = 50$, the interval is not symmetric about the point estimate; asymmetric intervals are the norm when the variable has a restricted range, as in the cases of correlations and proportions.

If an author elects to use the conventional confidence level of 95%, this should be stated the first time a CI appears in a paragraph, and the simple bracket format used thereafter to signal that the values in the brackets are the lower and upper limits respectively of the 95% CI. We recommend general use of 95% CIs, for consistency and to assist interpretation by readers, but particular traditions or special circumstances may justify choice of 99%, 90%, or some other CIs. If an author elects to use CIs with a different level of confidence, then that should be stated in every case throughout the manuscript, for example: "The mean improvement was 1.20 scale points, 90% CI [–0.40, 2.80]."

In a table, 95% CIs may similarly be reported as two values in square brackets immediately following the point estimate. Alternatively, the lower and upper limits of the CIs may be shown in separate labeled columns.

Altman, Machin, Bryant, and Gardner (2000) explained how to calculate CIs for a range of variables widely used in medicine, and provided software to assist. The variables covered include correlations, proportions, odds ratios, and regression coefficients. Cumming and Finch (2001) explained how to calculate CIs for Cohen's *d*. Grissom and Kim (2005) also provided advice on how to calculate CIs for many measures of ES.

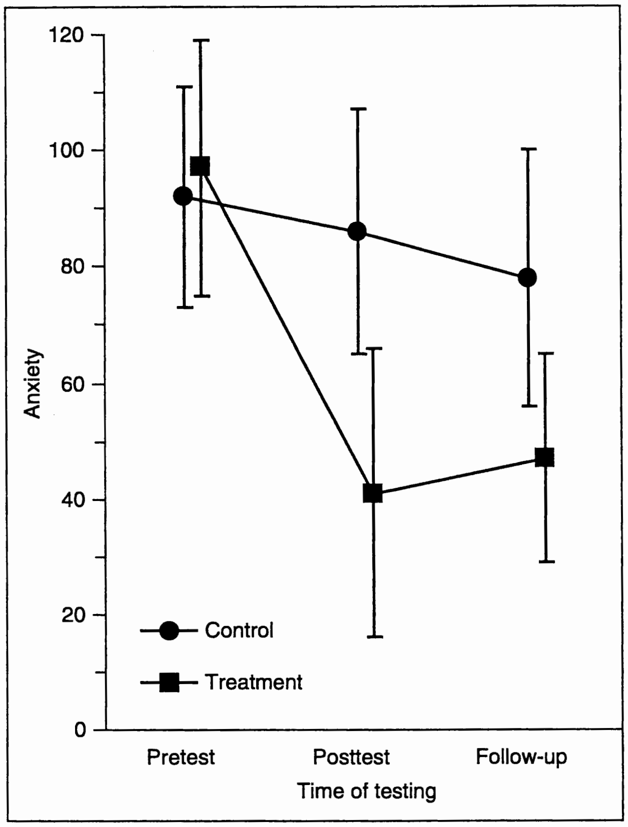
## 7. Figures with CI Error Bars

Wilkinson et al. (1999) advised that authors should: "In all figures, include graphical representations of interval estimates whenever possible" (p. 601). We agree, and Cumming and Finch (2005) discussed the presentation and interpretation of error bars in figures. A serious problem is that the familiar graphic used to display error bars in a figure, as shown in Figure 7.2, can have a number of meanings. The extent of the bars could indicate SD, standard error (SE), a 95% CI, a CI with some other level of confidence, or even some other measure of some variability. Cumming, Fidler, and Vaux (2007) described and discussed several of these possibilities. The most basic requirement is that any figure with error bars must include a clear statement of what the error bars represent. A reader can make no sense of error bars without being fully confident of what they show, for example 95% CIs, rather than SDs or SEs.

CIs are interval estimates and thus provide inferential information about the ES of interest. CIs are therefore almost always the intervals of choice. In medicine it is CIs that are recommended and routinely reported. In some research fields, however, including behavioral neuroscience, SE bars (error bars that extend one SE below and one SE above a mean) are often shown in figures. Unless sample size is less than about 10, SE bars are about half the width of the 95% CI, so it is easy to translate visually between the two. But SE bars are not accurately and directly inferential intervals, so CIs should almost always be preferred.

Cumming et al. (2007) found that, in leading psychology journals, from 1998 to 2006 there was an increase from 11% to 38% of articles that included at least one figure with error bars. That is a dramatic and welcome increase. However, even recently, 47% of those articles showed SE bars, and 34% did not make clear what the error bars represented. It is encouraging that many more authors are now including error bars in figures, but it remains a major problem that they often do not appreciate the desirability of using CIs, and the critical importance of making clear in every case what the error bars represent.

Figure 7.2 shows means with CIs for a hypothetical two-group experiment with a repeated measure. A treatment group was compared with a control group, and three applications of an anxiety scale provided pre-test, post-test, and follow-up measures. The figure illustrates several important issues. First, a knowledgeable practitioner might feel that the CIs are surprisingly and discouragingly wide, despite the reasonable group sizes ($N = 23$ and 26). It is an unfortunate reality across the social sciences that error variation is usually large. CI width represents accurately the uncertainty inherent in a set of data, and we should not shoot the messenger by being critical of CIs themselves for being too wide. The



*Note*: Mean anxiety scores and 95% CIs for a fictitious study comparing a Treatment (*n* = 23) and a Control (*n* = 26) group, at each of three time points: pre-test, post-test, and follow-up. Means have been displaced slightly so all CIs can be clearly seen.

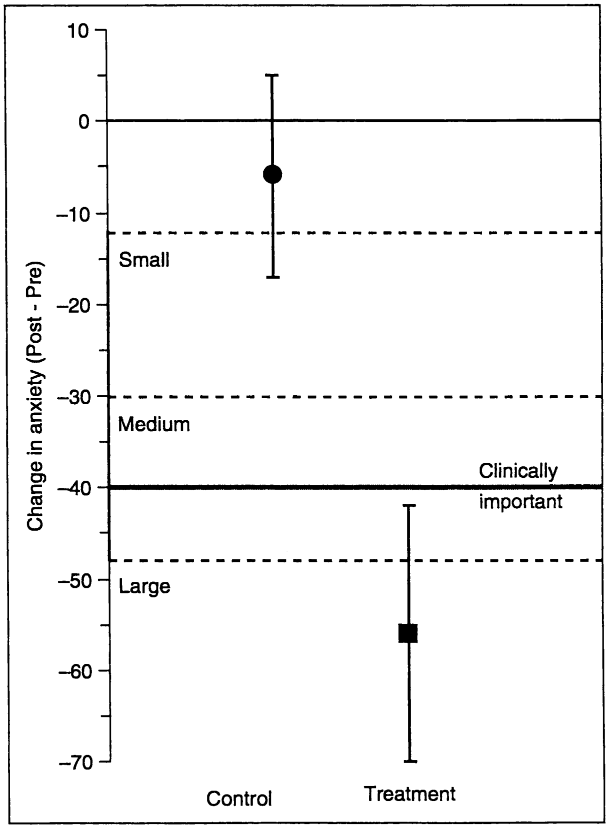
**Figure 7.2** Error Bar Display.

problem is NHST, with its simplistic reject or don't reject outcome, which may delude us into a false sense of certainty, when in fact much uncertainty remains. Cohen (1994) said, "I suspect that the main reason they [CIs] are not reported is that they are so embarrassingly large!" (p. 1002). We should respond to the message of large error variation by making every effort to improve experimental design and use larger samples, but must acknowledge the true extent of uncertainty by reporting CIs wherever possible.

Cumming and Finch (2005) provided rules of eye to assist interpretation of figures such as Figure 7.2. For means of two independent groups, the extent of overlap of the two 95% CIs gives a quick visual indication of the approximate $p$ value for a comparison of the means. If the intervals overlap by no more than about half the average of the two MOEs, then $p < .05$. If the intervals have zero overlap—the intervals touch end-to-end—or there is a gap between the intervals, then $p < .01$. In Figure 7.2, the control and treatment means at pre-test, for example, overlap extensively, and so $p$ is considerably greater than .05. At post-test, however, the intervals have only a tiny overlap, so at this time point $p$ for the treatment vs. control comparison is approximately .01. At follow-up, overlap is about half the length of the average of the two overlapping arms (the two MOEs), and so $p$ is approximately .05.

It is legitimate to consider overlap when the CIs are on independent means, but when two means are paired or matched, or represent a repeated measure, overlap of intervals is *irrelevant* to the comparison of means, and may be misleading. Further information is required, namely the correlation between the two measures, or the SD of the *differences*. For this reason it is not possible to assess in Figure 7.2 the *p* value for any within-group comparison, such as the pre-test to post-test change for the treatment group. Belia, Fidler, Williams, and Cumming (2005) reported evidence that few researchers appreciate the importance of the distinction between independent and dependent means when interpreting error bars. If CIs in figures are to be used to inform the interpretation of data—as we advocate—it is vital that figures make very clear the status of each independent variable. For between-subject variation, or independent means, intervals can be directly compared. For within-subject variation, a repeated measure, or dependent means, intervals may not be compared.

This important issue was illustrated further by Cumming and Finch (2005). If it seems puzzling, think of it in terms of the two familiar *t* tests: For two independent means, the independent *t* test is based on variation within the two groups; this variation determines the two CIs, and so those CIs are informative about the comparison of the two means. For paired data, by contrast, the paired *t* test is based on the variation in the paired differences, which is often in practice much smaller than the variation within either group. The variation in the differences is not represented by the CIs on the two means, and so these intervals are irrelevant to assessment of the mean difference—which can, however, be assessed if we have the CI on the difference, as shown in Figure 7.3 for the difference in each group between pre-test and post-test.



*Note*: Mean change in anxiety score from pre-test to post-test, for Treatment and Control groups, with 95% CIs, for the data shown in Figure 2. Dotted lines indicate reference values for changes considered small, medium, and large, and the grey line indicates the change considered to be clinically important.

**Figure 7.3** Confidence Intervals for Mean Differences.

ȝ a
ne
:ed

ay delude us into a false
"I suspect that the main
02). We should respond
rimental design and use
g CIs wherever possible.
f figures such as Figure
CIs gives a quick visual
intervals overlap by no
als have zero overlap—
< .01. In Figure 7.2, the
nd so *p* is considerably
at this time point *p* for
erlap is about half the
approximately .05.
s, but when two means
als is *irrelevant* to the
required, namely the

It is a problem that many current software packages do not sufficiently support the preparation of figures with error bars. In Figure 7.2, for example, the means are slightly offset horizontally so that all CIs can be seen clearly, but few packages make it easy to do this. One solution is to use Microsoft Excel. Figure 7.2 was prepared as an Excel scatterplot, which requires the horizontal and vertical coordinates for each point to be specified, so means can readily be displayed with a small horizontal offset.

In summary, the Results section should report point and interval estimates for the ESs of interest. Figures, with 95% CIs shown as error bars, should be presented wherever that would be informative. Every figure must make clear what error bars represent, and must describe the experimental design so a reader can understand whether each independent variable varies between or within subjects.

## 8. Interpretation of Effect Sizes

A primary purpose of the Discussion section is to present a substantive interpretation of the main ES estimates, and to draw out the implications. One unfortunate aspect of NHST is that the term *significant* is used with a technical meaning—a small *p* value was obtained—whereas in common language the word means "important." Kline (2004) recommended the word simply be dropped, so that if a null hypothesis is rejected we would say "a statistical difference was obtained." The common practice of saying "a significant difference was obtained" almost insists that a reader regard the difference as important, whereas it may easily be small and of trivial importance, despite yielding a small *p* value. Judging whether an ES is large or important is a key aspect of substantive interpretation, and requires specialist knowledge of the measure and the research context. We recommend that, if reporting NHST, either avoid the term "significant," as Kline recommended, or make its technical meaning clear by saying "statistically significant." When discussing the importance of a result, use words other than "significant," perhaps including "clinically important," "educationally important," or "practically important."

Cohen (1988, pp. 12–14) discussed the need for reference standards for the interpretation of common standardized and unit-free ES measures. He suggested standards that have become well known and widely used. For example, for Pearson correlation he suggested that values of .1, .3, and .5 can be regarded as small, medium, and large, respectively; and for Cohen's *d* he suggested similar use of .2, .5, and .8. An advantage of such standards is that ESs can be compared across different measures. He argued that the values and labels he chose are likely to be judged reasonable in many situations in the social sciences, but he stated clearly that they were arbitrary, and "were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible" (p. 532). Sometimes numerically tiny differences may have enormous theoretical importance, or indicate a life-saving treatment of great practical value. Conversely, a numerically large effect may be unsurprising and of little interest or use. Knowledgeable judgment is needed to interpret ESs (How large? How important?), and a Discussion section should give reasons to support the interpretations offered, and sufficient contextual information for a reader to come to an independent judgment.

## 9. Interpretation of Confidence Intervals

The correct way to understand the level of confidence, usually 95%, is in relation to indefinitely many replications of an experiment, all identical except that a new random sample is taken each time. If the 95% CI is calculated for each experiment, in the long run 95% of these intervals will include the population mean $\mu$, or other parameter being estimated. For our sample, or any particular sample, the interval either does or does not include $\mu$, so the probability that this particular interval includes $\mu$ is 0 or 1, although we will never know which. It is misleading to speak of a probability of .95, because that suggests the population parameter is a variable, whereas it is actually a fixed but unknown value.

ort the preparation of
horizontally so that all
to use Microsoft Excel.
nd vertical coordinates
orizontal offset.
for the ESs of interest.
would be informative.
experimental design so
within subjects.

retation of the main ES
is that the term *signifi-*
as in common language
be dropped, so that if a
" The common practice
regard the difference as
yielding a small *p* value.
rpretation, and requires
mend that, if reporting
e its technical meaning
a result, use words other
important," or "practi-

e interpretation of com-
ave become well known
es of .1, .3, and .5 can be
ested similar use of .2, .5,
different measures. He
n many situations in the
t forth throughout with
ble" (p. 532). Sometimes
or indicate a life-saving
y be unsurprising and of
How large? How impor-
ations offered, and suffi-
ent.

tion to indefinitely many
is taken each time. If the
vals will include the pop-
ny particular sample, the
lar interval includes $\mu$ is 0
ability of .95, because that
but unknown value.

Here follow some ways to think about and interpret a 95% CI (see also Cumming & Finch, 2005):

- The interval is one from an infinite set of intervals, 95% of which include $\mu$. If an interval does not contain $\mu$, it probably only just misses.
- The interval is a set of values that are *plausible* for $\mu$. Values outside the interval are relatively implausible—but not impossible—for $\mu$. (This interpretation may be the most practically useful.)
- We can be 95% confident that our interval contains $\mu$. If in a lifetime of research you calculate numerous 95% CIs in a wide variety of situations, overall, 95% of these intervals will include the parameters they estimate, and 5% will miss.
- Values around the center of the interval are the best bets for $\mu$, values towards the ends (the lower and upper limits) are less good bets, and values just outside the interval are even less good bets for $\mu$ (Cumming, 2007b).
- The lower limit is a likely lower bound of values for $\mu$, and the upper limit a likely upper bound.
- If the experiment is replicated, there is on average an 83.4% chance that the sample mean (the point estimate) from the replication experiment will fall within the 95% CI from the first experiment (Cumming, Williams, & Fidler, 2004). In other words, a 95% CI is approximately an 83% *prediction interval* for the next sample mean.
- The MOE is a measure of precision of the point estimate, and is the likely largest error of estimation, although larger errors are possible.
- If a null hypothesized value lies outside the interval, it can be rejected with a two-tailed test at the .05 level. If it lies within the interval, the corresponding null hypothesis cannot be rejected at the .05 level. The further outside the interval the null hypothesized value lies, the lower is the *p* value (Cumming, 2007b).

The last interpretation describes the link between CIs and NHST: Given a CI it is easy to note whether any null hypothesized value of interest would be rejected, given the data. Note, however, the number and variety of interpretations of a CI that make no reference to NHST. We hope these will become the predominant ways researchers think of CIs, as CIs replace NHST in many situations.

Authors may choose from the options above to guide their use of CIs to interpret their results. They may, for example, give substantive interpretation of the point estimate, and of the lower and upper limits of the CI, and thus cover the likely full range of plausible values for the parameter being estimated. Cumming et al. (2007) found that reporting of CIs in leading psychology journals increased from 4% to 11% of articles, between 1998 and 2005–2006. However, in only 24% of cases where CIs were reported were the intervals interpreted, or used explicitly to support data interpretation. As the *Publication Manual* recommends, "wherever possible, base discussion and interpretation of results on point and interval estimates" (APA, 2010, p. 34).

Figure 7.3 shows for the two groups the mean differences between pre-test and post-test, for the data presented in Figure 7.2. The figure includes 95% CIs on those differences, and there are reference lines that indicate the amounts of improvement judged by the researchers to be small, medium and large, and of clinical importance. These lines were created by adding further data series, with labels, to the Excel scatterplot used to present the mean differences with CIs. The CIs in Figure 7.3 allow us to conclude that, for the control group, the change from pre-test to post-test is around zero, or at most small; for the treatment group the change is of clinical importance, and likely to be large or even very large.

## 10. Meta-analytic Thinking

Figure 7.1 shows, as we mentioned earlier, the meta-analytic combination of hypothetical results from previous research. It also shows the point and interval estimates found in the current experiment, and a second meta-analysis that combines these with the earlier results. Considering that particular effect, our experiment has advanced the state of knowledge from the first to the second of the combined ES estimates, marked with square symbols in Figure 7.1. The Introduction and Discussion sections of the manuscript should both consider current research in the context of past results and likely future studies. This is meta-analytic thinking (Cumming & Finch, 2001), and it guides choice of what statistics are most valuable to report, and how results are interpreted and placed in context. Wilkinson et al. (1999) made several statements that outline what is needed:

> Reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. ... Reporting effect sizes also informs ... meta-analyses needed in future research. ... Collecting intervals across studies also helps in constructing plausible regions for population parameters (p. 599).
>
> ...
>
> Do not interpret a single study's results as having importance independent of the effects reported elsewhere in the relevant literature. ... The results in a single study are important primarily as one contribution to a mosaic of study effects (p. 602).

A forest plot (Figure 7.1) can display point and interval ES estimates expressed in any way—as original units, or in standardized form, or as some unit-free measure. For many types of research a forest plot can conveniently summarize current and past research in terms of estimation, and thus bring together all the ES, CI, and MA components that we have discussed in this chapter. It is important that authors, reviewers, and editors work together to help advance the social sciences as much as possible from the blinkered, dichotomous thinking of NHST to the richer and more informative research communication described in this chapter.

## Notes

[1] This research was supported by the Australian Research Council.

[2] Cumming and Finch (2001, 2005), Cumming, Williams, and Fidler (2004), and Cumming (2007b) are accompanied by components of ESCI ("ess-key"; Exploratory Software for Confidence Intervals), which runs under Microsoft Excel. Components of ESCI are available from www.latrobe.edu.au/psy/esci.

## References

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence* (2nd ed.). London: British Medical Journal.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cumming, G. (2007a). Confidence intervals. In G. Ritzer (Ed.) *The Blackwell encyclopedia of sociology* (Vol. II, pp. 656–659). Oxford, UK: Blackwell.

Cumming, G. (2007b). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics, 29,* 89–93.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

pothetical results from
rrent experiment, and
g that particular effect,
nd of the combined ES
scussion sections of the
and likely future stud-
ice of what statistics are
Wilkinson et al. (1999)

effects is essential to
d in future research.
gions for population

f the effects reported
tant primarily as one

ed in any way—as orig-
pes of research a forest
nation, and thus bring
ter. It is important that
ces as much as possible
informative research

(2007b) are accompanied
ich runs under Microsoft

d.). London: British Medical

gical *Association* (6th ed.).

intervals and standard error

um.

*ology* (Vol. II, pp. 656–659).

vels of confidence. *Teaching*

nin, N., & Wilson, S. (2007).

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology, 177*, 7–11.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 530–572.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170–180.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299–311.

Di Stefano, J., Fidler, F., & Cumming, G. (2005). Effect size estimates and confidence intervals: An alternative focus for the presentation and interpretation of ecological data. In A. R. Burk (Ed.), *New trends in ecology research* (pp. 71–102). Hauppauge, NY: Nova Science.

Fidler, F. & Cumming, G. (2007). The new stats: Attitudes for the twenty-first century. In J. W. Osborne (Ed.), *Best practice in quantitative methods* (pp. 1–12). Thousand Oaks, CA: Sage.

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics, 33*, 615–630.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83–105). Malden, MA: Blackwell.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Smithson, M. (2002). *Confidence intervals.* Thousand Oaks, CA: Sage.

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology: An International Review, 57*, 589–608.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist, 54*, 594–604.