Running head: EFFECT SIZES AND CONFIDENCE INTERVALS

Effect Sizes and Confidence Intervals

Fiona Fidler[1] and Geoff Cumming[2]

[1]The University of Melbourne, and [2]La Trobe University

Melbourne, Victoria, Australia

Contact author:

Geoff Cumming

School of Psychology and Public Health,

La Trobe University, Victoria, Australia 3086

Email: g.cumming@latrobe.edu.au

Fidler, F., & Cumming, G. (2019). Effect sizes and confidence intervals. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.) *The reviewer's guide to quantitative methods in the social sciences,* (2nd ed., Chapter 6) (pp. 72-85). London: Routledge.

An *effect size* 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 regression weight, a standardized measure of a difference, or the percentage of variance accounted for. Most research questions in the social sciences are best answered by finding estimated effect sizes, meaning *point estimates* of the true effect sizes in the population. Grissom and Kim (2012) provided a comprehensive discussion of effect sizes, and ways to calculate effect size estimates. Our discussion mainly focuses on experimental designs, but much of the discussion is relevant also to other types of research.

A *confidence interval* (CI), most commonly a 95% CI, is an *interval estimate* of a population effect size, which 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 length of the CI. The MoE is our measure of precision. Cumming and Finch (2005) provided an introduction to CIs and their advantages.

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 important and urgent that social scientists shift from relying mainly on NHST to using better techniques, including especially effect sizes, CIs, and meta-analysis. The changes needed are discussed in detail by Cumming (2012, 2014) and Kline (2013).

The most important current development in research methodology is the rise of Open Science, a set of practices designed to increase the openness, integrity, and reproducibility of research. One precursor was the classic article by Ioannidis (2005) that identified three problems which, together, suggested that “Most published research results are false” — to quote from the article’s title. The problems are (1) selective publication, especially of results that are statistically significant; (2) the imperative to achieve $p < .05$, which prompts

researchers to select and tweak what they report until they can claim statistical significance; and (3) the widespread belief that once a result has achieved $p < .05$ and been published it need not be replicated. Ioannidis identified over-reliance on NHST as the common factor underlying all three problems. Another key contribution was that of Simmons, Nelson, and Simonsohn (2011), who explained that the very large number of choices typically made by researchers as they analyzed their data and chose what to report meant that statistical significance could usually be found, no matter what the results. Expanding on the second problem of Ioannidis, they described many of those choices as *questionable research practices*, including for example, testing some extra participants after seeing the data, dropping some aberrant data points, and selecting which measures or comparisons to highlight. The use of questionable research practices to achieve $p < .05$ is called *p-hacking*, which is a pernicious and totally unacceptable practice. Widespread disquiet about *p* values and how they are used led the American Statistical Association to make a strong statement about their shortcomings (Wasserstein & Lazar, 2016).

The Center for Open Science (COS, cos.io) was founded in 2013 to promote Open Science, including especially the conduct of replications. It provides the Open Science Framework (OSF, osf.io), which is an invaluable and freely available online resource for researchers wishing to adopt Open Science practices. One fundamental Open Science practice is *preregistration* of a research plan, including a data analysis plan, in advance of running a study. Preregistering, then following the plan closely, helps ensure that planned and exploratory analysis can be clearly distinguished. It should also lead to reporting the study, whether in a journal or in an enduring accessible repository such as OSF, whatever results the study obtains. This would be a large step towards overcoming Ioannidis' first problem of selective publication.

Adopting estimation and meta-analytic perspectives helps us overcome the three Ioannidis problems, but more is required—questionable research practices such as the dropping of apparently aberrant results, or reporting only some selected comparisons, can be just as damaging when using CIs as with NHST. The full range of Open Science practices is needed.

Before assessing a manuscript, a reviewer should be familiar with the target journal's policies on Open Science issues, in particular the *Transparency and Openness Promotion (TOP) Guidelines*, which are available at cos.io/top/. In relation to various Open Science practices that we sketch below, such as preregistration, or provision of open data, a journal might encourage the practice, or even require it. It may offer badges, created by the Center for Open Science (tiny.cc/badges), to recognize articles that provide open data or open materials, or that preregistered the study being reported. Any manuscript needs to be assessed against such policies of the target journal.

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 statistical methods and Open Science.

Table 6.1. *Desiderata for Effect sizes and Confidence Intervals*

Desideratum	*Manuscript 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 design—whether experimental or otherwise—and	I, M

procedure is explained and justified in terms of appropriateness for obtaining precise estimates of the target effect sizes.	
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. Where possible a detailed research plan including data analysis plan was preregistered, then followed.	M
6. Results are presented and analyzed in terms of point estimates of the effect sizes.	R
7. The precision of effect size estimates is presented and analyzed in terms of confidence intervals.	R
8. Wherever possible, results are presented in figures, with confidence intervals.	R, D
9. Effect sizes are given substantive interpretation.	D
10. Confidence intervals are given substantive interpretation.	D
11. Meta-analytic thinking is used to interpret and discuss the findings. Replication is considered.	D
12. Where possible, full details of the materials and procedure are made openly available online.	D
13. Where possible, the data are made openly available online.	D

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

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 ± 0.02ºC, which signals that 27.35 is the point estimate of the boiling point, and 0.02 is the precision of that estimate. Correspondingly, it is most informative if the psychologist reports the effect of the psychotherapy as an effect size—the best estimate of the amount of change the therapy brings about—and a 95% CI to indicate the precision of that estimate. This approach can be contrasted with the impoverished dichotomous thinking (there is, or is not, an effect) that is prompted by NHST.

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

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

Expressions like these naturally lead to answers that are effect size estimates. Contrast these with statements like, “We investigated whether this treatment has an effect,” which suggests 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 effect size that answers these 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 betray 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 a statistically significant effect. That is an impoverished and 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.

Past research should, wherever possible, be discussed in terms of the point and interval estimates obtained for the effects of interest. Most simply, an effect size is a mean or other measurement in the original measurement units: The average extent of masked priming was 27ms; 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 effect size measure may be units-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 effect size measures indicate percentage of variance accounted for, such as R^2, as often reported in multiple regression, and η^2 or ω^2, as often reported with ANOVA. An important class of effect size measures are *standardized* effect sizes, including Cohen's *d* and Hedges' *g*. These 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*. The most appropriate effect size measure needs to be chosen for each research question, in the context of the research design. Grissom and Kim (2012) is an excellent source of assistance with the choice, calculation and presentation of a wide variety of effect size measures.

The introduction to the manuscript should focus on the effect size 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* allows that to be done quantitatively. Hunt (1997) gave a general introduction to meta-analysis, and an explanation of its importance. Borenstein, Hedges, Higgins, and Rothstein (2009), Cooper (2010), and Cumming (2012, Ch. 7-9) provided guidance for conducting a meta-analysis, and Chapter 19 of this volume discusses meta-analysis in more detail.

Figure 6.1 is a *forest plot*, which presents the results of 12 studies, and their combination by meta-analysis. The result of each study is shown as a point estimate, with its CI. The result of the meta-analysis 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 each empirical article to cite a meta-analysis —or, if none is available, wherever possible to carry out and report a new meta-analysis—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 in medicine, and should be used more widely.

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 an effect of a specified size, should this exist. The power approach was advocated by Jacob Cohen, and his book (Cohen, 1988) provided tables and advice (see also Chapter 26, this volume). An Internet search readily identifies freely-available software to carry out power calculations, including G*Power (tiny.cc/gpower3). 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 approach to selecting *N*.

An important advance in statistical practice is routine use of precision, meaning the MoE, in planning a study, as well as in discussion and interpretation of results. Cumming (2012, Ch. 13) described such a *precision for planning* 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 with a chosen target length: How large must *N* be for the expected 95% CI to be no longer than, for example, 60ms? Given a chosen experimental design, what sample size is needed for the expected MoE to be 0.2 units of Cohen's *d*? For two independent groups each of size *N*, Figure 6.2 shows a graph of required *N* against expected MoE, expressed in units of σ, the population SD.

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 Method 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—tend to be less common, but may be ethically problematic if they subject a needlessly large number of participants to an uncomfortable or time-consuming procedure. 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 effect sizes 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 effect sizes 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. Preregistration of a Research Plan

The OSF makes it easy for researchers or students to preregister a detailed research plan, including a data analysis plan, in an online repository where it is date-stamped and cannot be changed. It can be kept confidential, or at any time the researcher can make it open to all. Preregistration has long been required for drug trials, but only recently has it become recognized as important in the social sciences. Wagenmakers et al. (2012) explained the benefits and importance of preregistration. Authors are free to submit a study that had been preregistered to any journal, but, since 2014, *Psychological Science* has offered a badge for preregistered studies. Increasing numbers of journals are encouraging preregistration and offering the badge.

If a manuscript claims preregistration, a link must be provided to the plan that was lodged before data collection commenced, and the study must have been conducted and analyzed in accordance with that plan. Data exploration beyond the planned analysis may be acceptable, but any results found by exploration may easily be cherry picked, mere capitalization on chance, and are at best speculations, perhaps for further investigation.

6. Results: Effect Sizes

The main role of the Results section is to report the estimated effect sizes that are the primary outcomes of the research. We mentioned in Desideratum 2 the wide range of possible effect size 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 effect size 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 meta-analysis to have to transform effect size estimates reported in a variety of ways into some common

measure, as the basis for conducting the meta-analysis. In medicine, odds ratio or log odds ratio are frequently used as the common effect size measure, but in social science Cohen's *d*, or Pearson's *r* correlations are frequently chosen as the basis for meta-analysis.

Often it may be useful 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 comparison of results over different studies, and the conduct of future meta-analysis. 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 Cohen's *d* value 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 effect size 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 our 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 we are studying, 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 make clear what basis was chosen for standardization.

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 effect size 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).

7. Results: Confidence Intervals

Following the *Publication Manual* of the American Psychological Association (APA, 2010) we recommend the following style for reporting CIs in text:

At the first occurrence in a paragraph write: "The mean decrease was 34.5 ms [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.

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: “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.

Cumming and Finch (2001) and Cumming (2012, Ch. 11) explained how to calculate CIs for Cohen’s *d*. Grissom and Kim (2012) provided advice on how to calculate CIs for many measures of effect size. Calculation of CIs can be straightforward, or may best be accomplished using computer-intensive methods, such as bootstrapping. Helpful software is becoming increasingly available (Cumming, 2014, p. 25).

8. Figures with Confidence Interval Error Bars

Whenever possible, researchers should provide figures that include 95% CIs. 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 3, can have a number of meanings. 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 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 effect size 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. When sample size is at least about 10, SE bars are about half the length 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.

Figure 6.3 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 long, 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 length 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 long. The 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 3. 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 6.3 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 6.3 the p value for any within-group comparison, such as the pre-test to post-test change for the treatment group. Belia, Williams, Fidler, 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.

It is a problem that many current software packages do not sufficiently support the preparation of figures with error bars. In Figure 6.3, 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 6.3 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 effect sizes 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.

9. Interpretation of Effect Sizes

A primary purpose of the Discussion section is to present a substantive interpretation of the main effect size 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 (2013) 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 effect size 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 recommends, or make its technical meaning clear by saying "statistically significant". When discussing the importance of a result, use words other than "significant", perhaps including "notable", "clinically important", or "educationally important".

Cohen (1988, pp. 12-14) suggested reference values for the interpretation of some effect size measures. 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. However, he stated that his reference values 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 effect sizes (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.

10. Interpretation of Confidence Intervals

A manuscript should not only report CIs, but also use them to inform the interpretation and discussion of results. 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 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 μ, or other parameter being estimated. For our sample, or any particular sample, the interval either does or does not include μ, so the probability that this particular interval includes μ 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.

Here follow some ways to think about and interpret a 95% CI (see also Cumming, 2012; Cumming & Finch, 2005).

- The interval is one from an infinite set of intervals, 95% of which include μ. If an interval does not contain μ, it probably only just misses.
- The interval is a set of values that are *plausible* for μ. Values outside the interval are relatively implausible—but not impossible—for μ. (This interpretation may be the most practically useful.)

- We can be 95% confident that our interval contains μ. If in a lifetime of research you calculate numerous 95% CIs in a wide variety of situations, overall, around 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 μ, 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 μ (Cumming, 2007).
- The lower limit is a likely lower bound of values for μ, and the upper limit a likely upper bound.
- If the experiment is replicated, there is on average about an 83% 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, 2007).

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 any of the options above to guide their use of CIs to interpret their results. As the *Publication Manual* recommends, "wherever possible, base discussion and interpretation of results on point and interval estimates" (APA, 2010, p. 34).

Figure 6.4 shows for the two groups the mean differences between pre-test and post-test, for the data presented in Figure 6.3. 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. The CIs in Figure 6.4 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.

11. Meta-analytic Thinking and Replication

Figure 6.1 shows, as we mentioned earlier, the meta-analytic combination of results from 12 studies, which might be all the available previous research. 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 measures and statistics are most valuable to report, and how results are interpreted and placed in context. A forest plot (Figure 6.1) can display point and interval effect size estimates expressed in any way—as original units, or in standardized form, or as some units-free measure. For many types of research a forest plot can conveniently summarize current and past research in terms of estimation.

Replication is at the heart of science, even if in social science it has too often been neglected. All manuscript authors should consider replication, which is part of meta-analytic thinking. Even if they cannot themselves immediately run a replication, they should do all they can to assist any future replication of their work.

Replications, especially, should be judged by how well they are planned and conducted, and not by the results they obtain. Some journals are adopting the highly desirable policy of reviewing a study in advance of data collection. If the research question, and proposed design and methods, are all judged of a sufficiently high standard, the report is accepted in advance, subject only to the study being conducted as planned. Reviewers should support this enlightened policy, which should help overcome the first and third Ioannidis problems.

12. Open Materials

Providing fully detailed information about the procedure and materials used for a study is necessary for readers to understand fully what was done, and also to provide maximum assistance to any future replication efforts. The full details may need to be provided in an online supplement to a journal article, and/or in a permanent repository such as OSF. Including a protocol video, which shows how a study was actually run, can be highly valuable for anyone seeking to run a replication. In an increasing number of journals, provision of open materials, in full detail, may be acknowledged by award of the Open Materials badge.

13. Open Data

Providing open access to the full data set has many benefits: It makes meta-analysis easier, allows anyone to check for errors of analysis and interpretation, and makes it much easier for others to replicate the work. In addition, researchers can analyze the data in different ways, perhaps to address different research questions. Of course, researchers must not post sensitive or identifying information about their participants, and they need to be sure their participants have consented to anonymous data sharing, and that the intention to provide open access to data was part of the initial application for ethics approval.

Researchers might feel that, having made the enormous effort to collect their precious data, they want to be able to use it as part of future research, rather than release it immediately to other researchers. When there are good reasons, it can be acceptable to delay release of full data while the original researchers work further with it, but usually 12 months should be the maximum delay before data are made openly available. Again, the journal may offer an Open Data badge to acknowledge that the full data set is available from an open enduring repository.

Our conclusion is that 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, and make every effort to encourage Open Science.

References

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.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

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.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

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

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. Available for free download from tiny.cc/tnswhyhow

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. Available for free download from tiny.cc/inferencebyeye

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

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.

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

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. Retrieved from http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124. Available for free download from tiny.cc/mostfalse

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. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington DC: American Psychological Association.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

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.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, Han L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632-638.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose, *The American Statistician, 70,* 129-133, doi: 10.1080/00031305.2016.1154108

Author Notes

Fiona Fidler is an Australian Research Council Future Fellow and Associate Professor in the School of BioSciences and the School of Historical and Philosophical Studies at The University of Melbourne. She is interested in how scientists and experts make decisions, and how methodological practices in science change over time. https://fionaresearch.wordpress.com/about/

Geoff Cumming (g.cumming@latrobe.edu.au) is Emeritus Professor, School of Psychology and Public Health, La Trobe University, Melbourne, Australia. His main research area is statistical cognition, and he encourages adoption of Open Science practices and use of estimation and meta-analysis ('the new statistics'). His first book focuses on the new statistics, and the second, an introductory textbook co-authored with Robert Calin-Jageman, focuses on Open Science as well as the new statistics. Both make extensive use of ESCI (Exploratory Software for Confidence Intervals), which runs under Microsoft Excel. For more information, and free download of ESCI, see: www.thenewstatistics.com

(Figure 6.1)

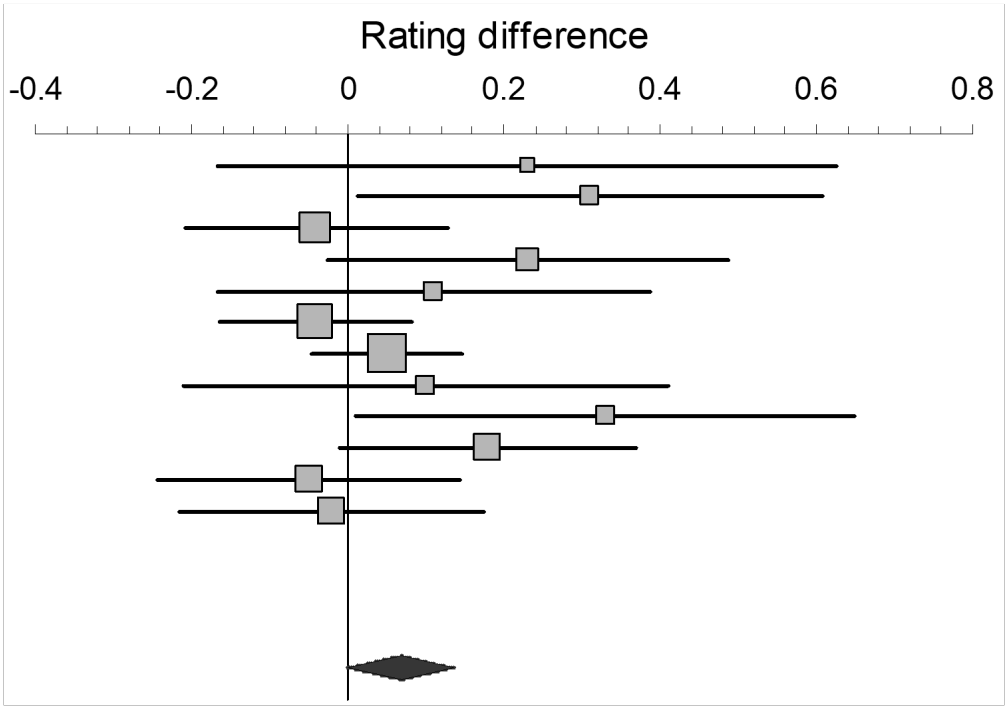

Figure 6.1. A forest plot, showing the results of 12 studies that estimated the mean rating difference between two types of stimuli. Squares indicate point estimates, and error bars the 95% CIs. The sizes of the squares indicate approximately the weights of the different studies in the meta-analysis. The result of the meta-analysis is displayed as a diamond, whose horizontal extent is the 95% CI.

(Figure 6.2)

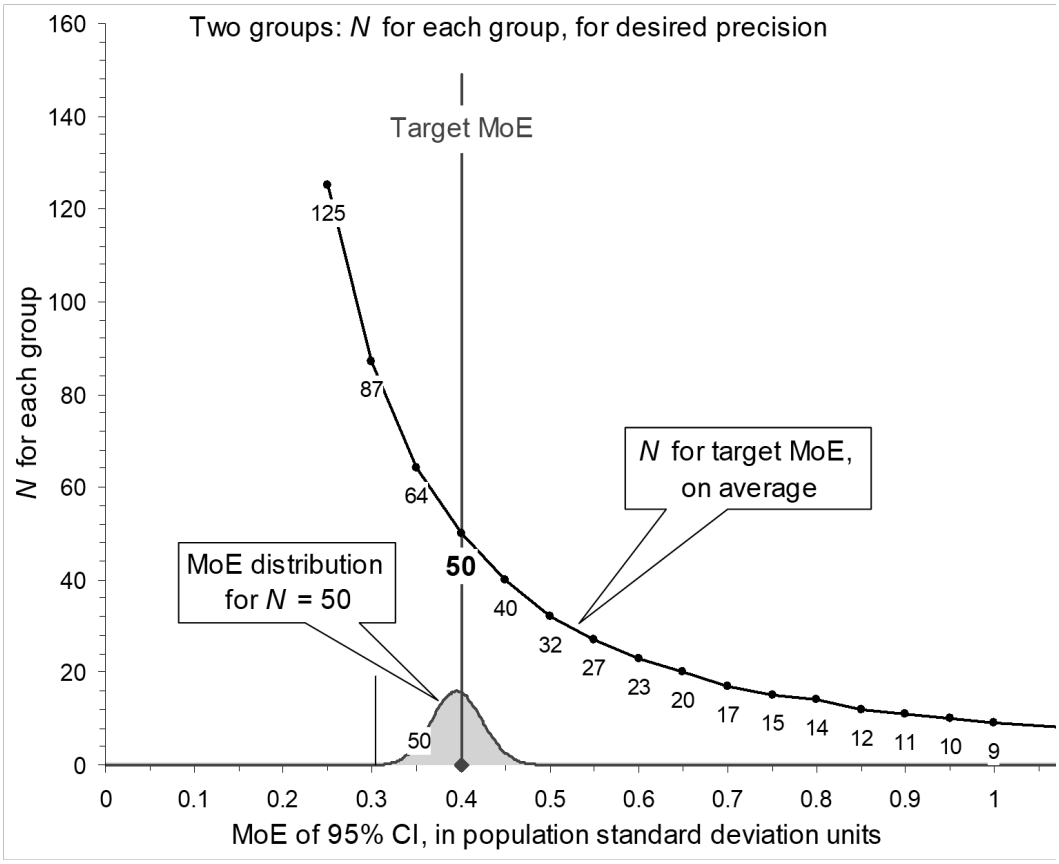

Figure 6.2. Graph required for precision for planning. The curve indicates the *N* required by a two independent groups study, each group of size *N*, for MoE of the 95% CI on the difference between the group means to be as shown on the horizontal axis, on average. The vertical cursor marks MoE = 0.4, in units of σ, the population SD. The shaded curve is the distribution of MoE values over a very large set of studies all having $N = 50$.

(Figure 6.3)

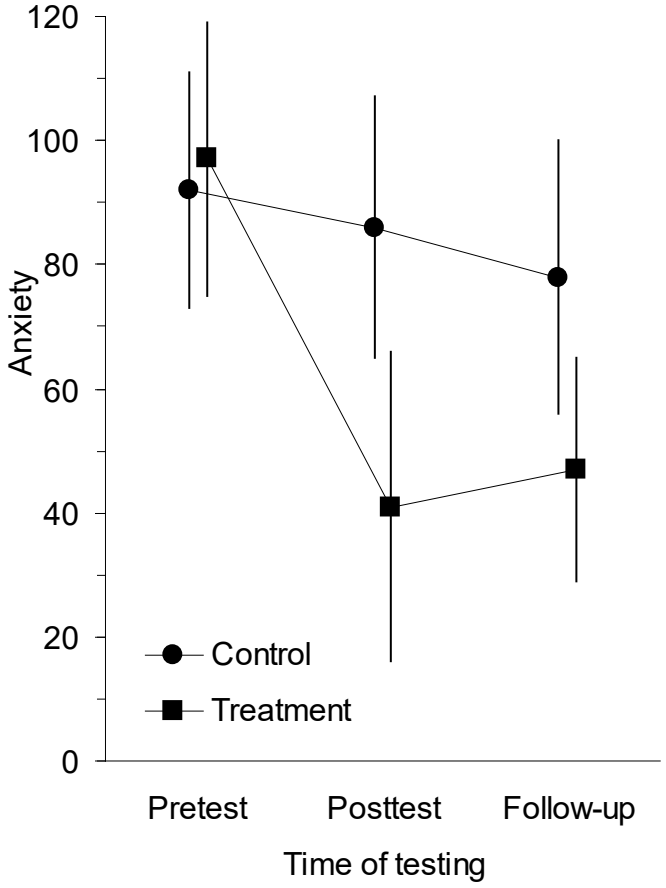

Figure 6.3. Mean anxiety scores and 95% confidence intervals (CIs) for a fictitious study comparing a Treatment ($N = 23$) and a Control ($N = 26$) group, at each of three time points: pre-test, post-test, and follow-up. Means have been displaced slightly so all CIs can be clearly seen.

(Figure 6.4)

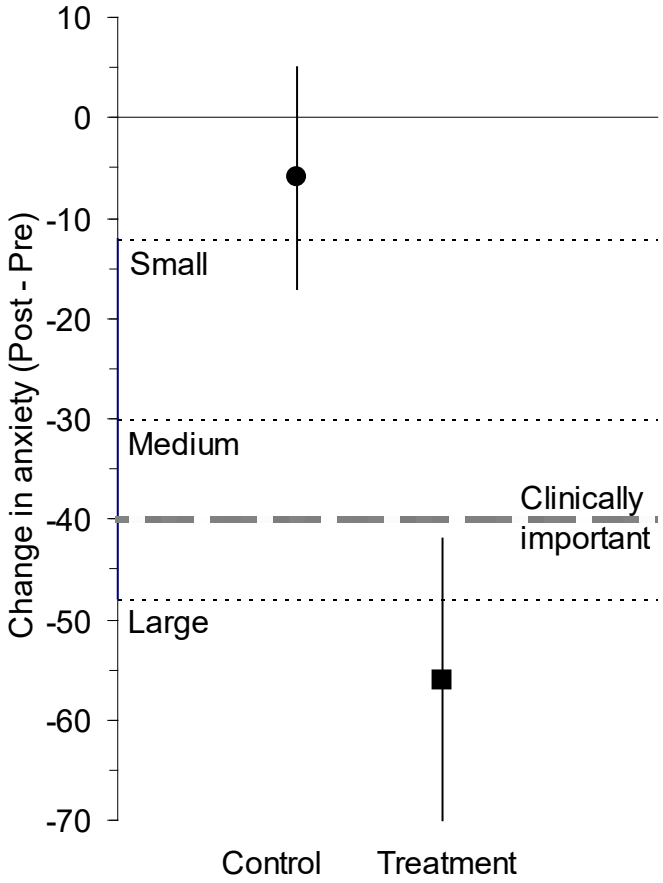

Figure 6.4. Mean change in anxiety score from pre-test to post-test, for Treatment and Control groups, with 95% CIs, for the data shown in Figure 6.3. Dotted lines indicate reference values for changes considered small, medium and large, and the grey line the change considered clinically important.