Australian
Psychologist

The New Statistics: A How-To Guide

Geoff Cumming

Statistical Cognition Laboratory, School of Psychological Science, La Trobe University, Melbourne

Estimation, based on effect sizes (ESs) and confidence intervals (CIs), is much better than null hypothesis significance testing (NHST). I refer to estimation and meta-analysis—which is the extension of estimation to multiple studies—as *the new statistics*. The techniques themselves are not new, but using them would for many psychologists be new, and a highly beneficial advance. I describe a six-step strategy for estimation, which starts with the statement of research questions in terms of "how large?" questions, rather than the dichotomous hypotheses of NHST. I outline how to use estimation to analyse the two-independent-groups and paired designs, and randomised control trials. I discuss the ES measures Cohen's *d* and correlation, *r*, in relation to estimation. I describe how to interpret results published using NHST by visualising the corresponding CIs, and give guidance for adopting the new statistics, even in a world that often still expects NHST. I emphasise the value of figures that display CIs, and describe freely available software running under Microsoft Excel that assists calculation of CIs and preparation of figures.

Key words: confidence intervals; data analysis; effect sizes; estimation; statistical reform; the new statistics.

My aim in this article is to outline in simple terms how to follow the unequivocal advice of the *American Psychological Association Publication Manual* (APA, 2010, p. 34) to base the interpretation of data on point and interval estimates wherever possible. I will start with a made-up example. Consider a randomised control trial (RCT) of a new form of psychotherapy for depression. The main finding might be reported using null hypothesis significance testing (NHST):

> ***NHST format.*** *"The new therapy demonstrated a statistically significant advantage over the control procedure,* p < *.01." A figure shows the pretest and posttest mean Beck Depression Inventory (BDI-II; Beck, Steer, Ball, & Ranieri, 1996) scores for the treatment and control groups. The figure suggests that the control group improved by an average of about 6 points and the treatment group about 11 points. The discussion and abstract emphasise that the finding is highly statistically significant,* p < *.01.*

Alternatively, you might read:

> ***Estimation format.*** *"The advantage of the new therapy over the old was an average 5.0 points on the BDI-II, 95% CI [1.4, 8.6]." The discussion and abstract focus on the 5-point advantage, the uncertainty of that result as indicated by the confidence interval (CI), and the clinical importance of a 5-point advantage.*

The *APA Manual* refers to point and interval estimates. A *point estimate* is simply our best estimate of the effect we are studying, often a mean or a difference between two means; it is also called the *effect size* (ES). So 5.0 points is our ES estimate. An *interval estimate* is a CI that quantifies the uncertainty in our ES estimate: A short CI is good news—we have relatively precise ES knowledge—but a long CI indicates greater uncertainty. Our CI of [1.4, 8.6] indicates that the true advantage of the new therapy is, most likely, somewhere between about 1 and 9 points. Drawing conclusions from data by using point and interval estimates is called *estimation*, and so the *APA Manual*'s statement tells us to use estimation to draw conclusions from our research results.

I use the term *the new statistics* to refer to estimation, meaning ESs and CIs, and also to meta-analysis—which is the extension of estimation to more than one study. The techniques themselves are not new, but using them would, for many researchers, be quite new, and a very beneficial change.

Most psychologists would find the NHST format earlier familiar and reassuring—"highly statistically significant" seems close to a statement of certainty. The NHST format has been, of course, overwhelmingly the dominant way psychologists have reported their research for more than half a century. By contrast, the estimation format is probably unfamiliar and perhaps disappointing because the CI is so long. The two formats are, however, reporting exactly the same results, so any feeling of reassurance or disappointment should be the same for the two.

The next section summarises briefly why estimation is better than NHST, and why it is so important for psychology to shift to the new statistics. Then I will introduce a six-step estimation strategy, and say more about ESs, CIs, and good ways to think about CIs. I will then illustrate the estimation approach to presenting data for several simple designs, and discuss briefly two ES measures: Cohen's *d* and correlation, *r*. We need, of course, to be able to read the NHST-based research literature, so I describe how to translate *p*-value results into an estimation format, for better understanding. Finally, I discuss the practical challenges of using the new statistics in a world that still, in many cases, expects *p*-values. This article is a companion to the

Correspondence: Geoff Cumming, Statistical Cognition Laboratory, School of Psychological Science, La Trobe University, Melbourne, Vic. 3086 Australia. email: g.cumming@latrobe.edu.au

This research was supported by the Australian Research Council. I thank Jacenta Abbott and Laura Anderson for valuable comments on a draft.

Accepted for publication 11 November 2012

doi:10.1111/ap.12018

brief article in *InPsych* (Cumming, 2012a) and to Cumming, Fidler, Kalinowski, and Lai (2012), which discusses the *Manual*'s statistical advice in more detail.

The New Statistics: Why?

There are positive and negative reasons why psychology and other disciplines should adopt the new statistics, and avoid NHST wherever possible. The main positive reason is simply that estimation is more informative, and gives direct, quantitative answers to our research questions. Estimation answers the question "How large is the effect?", which is of most relevance to practitioners as well as researchers. The answer "around 5, most likely between 1 and 9" allows us to use our professional judgement about the clinical importance of such an average decrease in Beck scores. An individual client may have a different experience, but that answer is our best guide to the average improvement the new therapy can offer.

The negatives are the deep flaws of NHST and the damage it has done to research progress, as described by many leading scholars. One central problem is that NHST seeks to answer the question "Is there an effect?", which is an impoverished question, suggesting our research need only provide a yes or no conclusion. Effects, however, come in every size from zero to tiny to small, and all the way up to enormous. NHST divides them arbitrarily into just two categories—"not significant" and "statistically significant"—which is misleading because the cutoff between the categories depends as much on our experimental design and sample size, N, as on the ES. A second crucial problem for NHST is that, if we repeat our experiment, we are likely to get a *very* different value of p. So a p-value gives us hardly any information at all. For a demonstration of the unreliability of p, see the video at tiny.cc/dancepvals. A declaration of "highly significant" seems to offer seductive certainty, but that is an illusion. The CI may be disappointingly long, but it tells us accurately about the uncertainty in our data. It is an intriguing question why psychology and other disciplines are so addicted to NHST—perhaps the seductive certainty—but it is only tradition and inertia that keeps it going, with almost no scholars defending it. All round, NHST is a terrible idea, and the sooner psychology moves to the new statistics, or other preferred techniques, the better.

The best summary of NHST problems and estimation advantages was given by Rex Kline (2004, Chapter 3; tiny.cc/klinechap3). I gave an introduction to the new statistics intended for general audiences in a radio broadcast (transcript and podcast at tiny.cc/geofftalk) and magazine article: tiny.cc/GeoffConversation. There is more in my book (Cumming, 2012b) and at the book's website: www.thenewstatistics.com.

That website also supports free download of ESCI ("ESS-key", *Exploratory Software for Confidence Intervals*), which runs under Microsoft Excel. ESCI is designed to illustrate many estimation concepts and to calculate and display CIs in simple situations. Most figures later come from ESCI—the captions state which ESCI page I used to make the figure. You can use those pages to explore the concepts illustrated and, in many cases, to prepare a similar figure for you own data. You can use any of Excel's capabilities to modify or extend the figure as you wish. Cumming (2012b, Appendix A) provides more assistance.

The Six-Step Estimation Strategy

Here is a basic strategy for using the new statistics:

1. ***Formulate research questions in estimation terms.*** Ask "How large is the effect?", "How many?", or "To what extent?". Avoid the dichotomous statements or questions prompted by NHST, such as "Test the hypothesis that there's no difference", or "Is the new therapy better?"
2. ***Identify the ESs that best answer the research questions.*** In our example, we focus on the difference between the mean change scores in the two groups.
3. ***From the data, calculate point and interval estimates (CIs) for these ESs.*** We found the mean advantage to be 5.0 points, 95% CI [1.4, 8.6]. (That is the format specified by the *APA Manual* [APA, 2010, p. 117] for reporting a CI, and what psychologists should routinely use.)
4. ***Make a figure, including CIs.*** There are examples later.
5. ***Interpret the ESs and CIs.*** Discuss the theoretical, clinical, and/or practical implications, in accord with the aims of the research.
6. ***Use meta-analysis where appropriate.*** Meta-analysis is a set of techniques for combining the results of a number of studies on related issues.

These steps may seem simplistic or obvious, but each differs from what has been common practice in psychology. Step 1 may in practice require a big change in thinking by many researchers, but may be the key to adopting the new statistics, because asking "How much?" naturally prompts a quantitative answer—an ES. Step 2 may seem trite, but can lead us to appreciate that we are interested in a difference, so need to calculate that difference, and also the CI on that difference.

ESs and CIs: The Basics

An ES is the amount of anything of research interest. We are all familiar with ESs, even if not by that name. A mean, the difference between two means, a proportion or percentage, a correlation, an odds ratio, a regression slope: These are all ESs. Some writers give the impression that only more complex measures, such as Cohen's d (see later) and η^2 (eta-squared) qualify as ESs. Those are indeed ESs, but so are many more familiar measures. A p-value, however, is *not* an ES. Usually, we use an ES estimate based on our data (e.g., the sample mean, M) as our point estimate of the corresponding population ES (the population mean, μ).

ES measures can be grouped in various ways. An ES might be expressed in original units, such as Beck points or centimetres or milliseconds, or transformed into a standardised form as a Cohen's d-value. It can often be useful to report both forms. Some ESs are units-free, such as correlation, r, and proportions. Some are expressed in a squared metric, such as R^2, a proportion of variance. Grissom and Kim (2012) provided advanced advice on ESs.

A 95% CI is designed so that, with the usual assumptions of random sampling from a normally distributed population, the interval will include μ, the population parameter we are

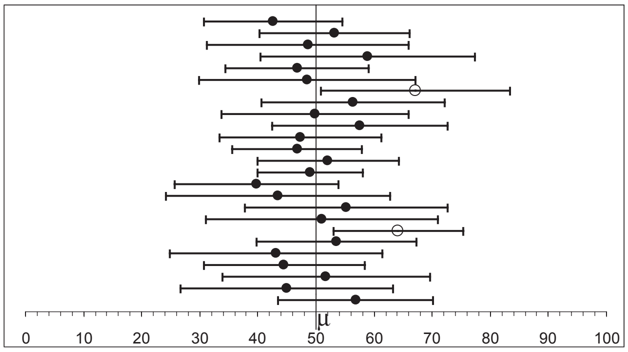

Figure 1 The dance of the confidence intervals (CIs): results of 25 replications of a simulated experiment, each comprising a single sample of $N = 30$ scores, from a normally distributed population with mean $\mu = 50$ (marked with a vertical line) and standard deviation $\sigma = 40$. Means (black dots) and 95% CIs are shown. The two means whose CI does not capture μ are shown as open circles—in Exploratory Software for Confidence Intervals (ESCI), they are red. Figure from the **CIjumping** page of **ESCI chapters 1–4**.

estimating, for 95% of an infinite sequence of replications of our experiment, as Figure 1 illustrates. Our single CI, calculated from our data, may or may not include μ, but we never know which. In a lifetime of seeing numerous 95% CIs, about 95% will include the population parameter and about 5% will miss.

I am focusing on 95% CIs, rather than 99% or any other level of confidence because they are most common. It is best to be consistent and always use 95% CIs, unless there are strong reasons to do otherwise. From now on, I will assume all CIs are 95% CIs.

Our CI, for example [1.4, 8.6], is randomly chosen from a sequence like that in Figure 1—plus the infinity of cases the figure could have illustrated. I call the sequence the *dance of the CIs* because, in ESCI, you can watch them dancing down the screen. We have little option but to interpret the single interval that we have, but we should always bear in mind that it is one from the dance. Figure 1 includes two CIs that just miss μ; in ESCI these are red, so our slogan for any single CI is that "it might be red". In real life, unfortunately, CIs are not coloured!

We can interpret our interval by saying that values around 5, the centre of our CI, are the most likely for μ, or our best bets for μ, but any values between 1.4 and 8.6 are plausible. Values outside the interval are less plausible—but not impossible, because our CI might be red. Cumming & Finch, (2005; tiny.cc/inferencebyeye) provided a fuller introduction to CIs.

Three Simple Designs

I will consider two very simple designs: Two independent groups, and the paired design. These may be familiar as the two "*t*-test designs". Then I will put them together and extend them slightly to form an RCT.

Two Independent Groups

Figure 2 shows the pretest means in our example, for the treatment and control groups, with their CIs. Showing two independent means with their error bars may be familiar, but note carefully what the error bars represent. It is extremely unfortunate that the error bar graphic can be used to display not only CIs, but also SE bars or even other quantities. It is essential that any figure with error bars states clearly what the bars represent. Cumming and Finch (2005) explained why CIs should always be preferred to SE bars, and Cumming, Fidler, and Vaux (2007; tiny.cc/errorbars101) discussed a range of issues relating to error bars.

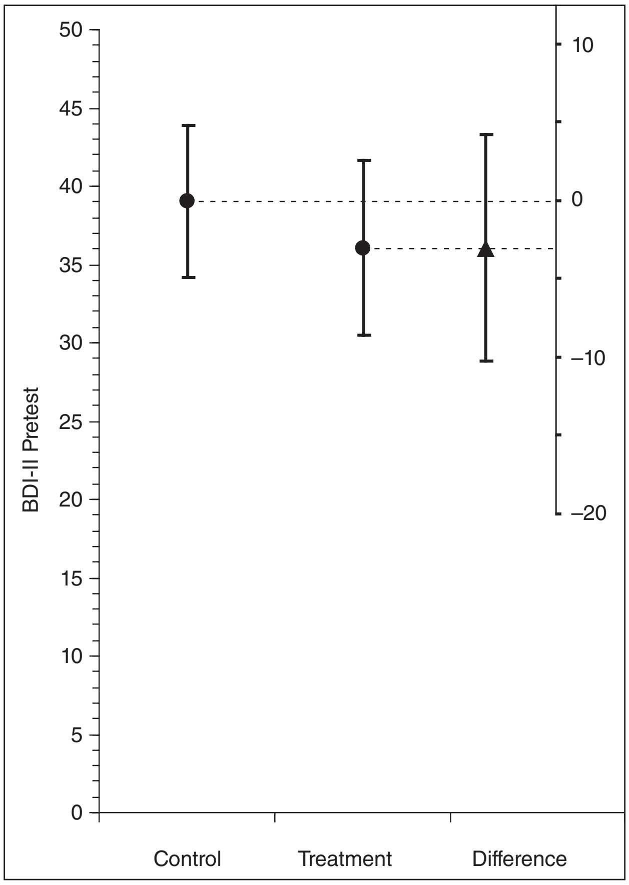

Figure 2 A fictitious two-independent-groups example. Pretest Beck Depression Inventory (BDI-II) means are shown for the control and treatment groups, with 95% confidence intervals (CIs). The difference between the means is marked by the triangle, with its 95% CI, on a floating difference axis at right. Figure from the **Data two** page of Exploratory Software for Confidence Intervals (**ESCI**) **chapters 5–6**.

Figure 2 includes a floating difference axis, a novel feature to display the ES, and its CI that are needed to answer our research question, which here asks about the difference at pretest between the two groups. The difference axis is displayed at right, its zero lined up with the control mean. The difference between the means is marked by the triangle: The treatment pretest mean was three points lower than that of the control group. The CI on the difference is longer than the CIs on the two group means because uncertainty in the difference is a compounding of the uncertainties in the two group means. The *margin of error* (MOE) is the length of one arm of a CI. MOE was 5.6 for the treatment CI, 4.9 for the control CI, and 7.3 for the difference CI. The pattern illustrated in Figure 2 is typical for two independent groups: The MOE (and therefore the CI) for the difference between the group means is usually about 40% longer than the MOEs (and the CIs) for the individual means.

Focus on the difference and its CI to conclude that, at pretest, the average difference was most likely between about −10 and 4. A CI can be used to carry out NHST: If the null hypothesised value lies within a 95% CI, the null hypothesis is not rejected at the .05, two-tailed level. If the interval does *not* include the null hypothesised value, reject. In this case, zero difference is well within the CI on the difference, so we can conclude there was no statistically significant difference between the groups at pretest. Be very careful, however, to avoid the common error of assuming we can therefore conclude there was no difference. It is much better to avoid NHST and to inspect and discuss the CI because it indicates the plausible range of true pretest differences, which is often, as in this case, quite long. Only if all values in the CI can be regarded as negligible—we have a very short CI close to zero—are we justified in concluding there was negligible average pretest difference between the groups.

If we are given a figure that shows only the separate group means with their CIs, in the independent-groups case, we can use the two separate CIs to assess the difference. Best is to visualise as in Figure 2 the floating difference axis, the difference, and the CI on the difference—seeing this interval as about 40% longer than the separate CIs. We could also note whether the two CIs overlap, touch, or are separated. Just touching indicates reasonable evidence of a difference, and is usually equivalent to a *p*-value of about .01 for the comparison of the two means. As usual, invoking NHST is not the best strategy, but with CIs on independent means it is possible. Cumming and Finch (2005) and Cumming (2012b, Chapter 6) have more.

Paired Data

Figure 3 shows the pretest and posttest means for the treatment group in our example. The improvement for that group was 11 points [8.4, 13.6], and that difference is shown, with its CI, on the floating difference axis. The striking contrast here is that the CI on the difference was so short, with MOE of 2.6, less than half that of the CI on either pretest or posttest. This short CI reflects, of course, the repeated measure—the pretest and posttest scores for a single group—and signals the sensitivity of the design. The correlation of the pretest and posttest scores in the treatment group was .89, which is not atypical for paired data. The short CI on the change score indicates we are estimating the change with reasonably high precision.

It is vital to understand the contrast between the independent and paired designs, and between Figures 2 and 3. With independent means (Figure 2), the CI on the difference is based on the same variability within groups as the two separate CIs. We can assess the difference by considering the CI on the difference, or the configuration of the two means and their CIs—do they overlap, touch, or miss? For paired data and Figure 3, however, the situation is quite different because the CI on the difference is based not on the variability within groups, but on the standard deviation (SD) of paired differences. The width of this CI reflects the correlation between the two measures (here, pretest and posttest), with a higher correlation giving a shorter CI on the difference. To assess the difference, we must have the CI on the difference; the CIs on the separate measures are irrelevant because they do not reflect the correlation. If a figure reporting repeated-measure results shows only the separate measures, with their CIs, we have no way of assessing the difference. In Figure 2, the floating difference axis is optional—if not provided, we can imagine it—but in Figure 3, it is essential, and without it, we cannot interpret the result.

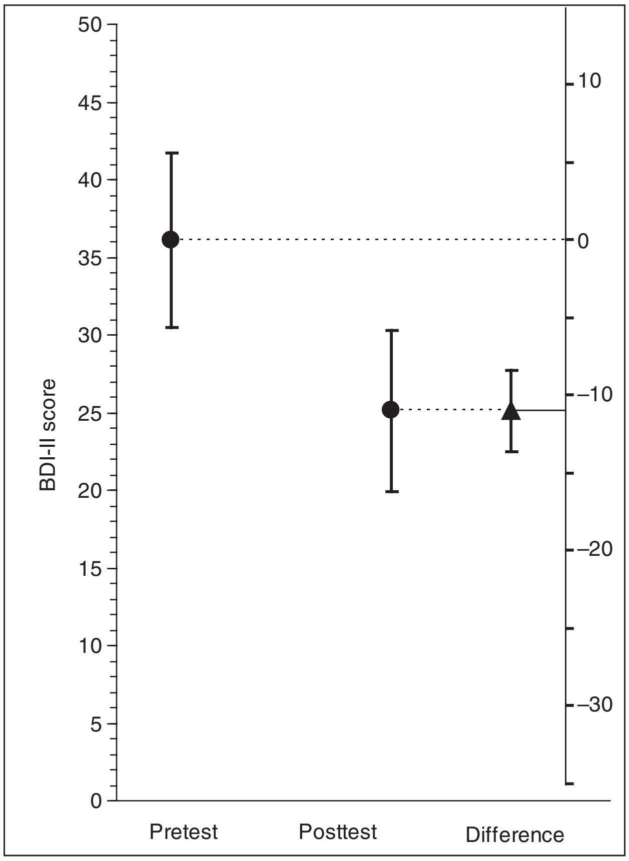

Figure 3 A fictitious paired data example. Beck Depression Inventory (BDI-II) pretest and posttest means for the treatment group are shown, with 95% confidence intervals (CIs). The mean of the paired differences is marked by the triangle, with its 95% CI, on a floating difference axis at right. Figure from the **Data paired** page of Exploratory Software for Confidence Intervals (**ESCI**) **chapters 5–6**.

This contrast between the two designs is crucial, so I will say a little more about it. It parallels the distinction between the independent-groups *t*-test and paired *t*-test. Independent *t* is calculated using the *SDs* within the two groups—the same *SDs* we use to calculate the CIs on the separate means in Figure 2. The CI on the difference in Figure 2 is based on these same within-group *SDs*. Paired *t*, however, is based on the *SD* of the paired differences—just as we use to calculate the CI on the difference in Figure 3. The *SDs* for the separate measures play no direct role in paired *t*, just as the CIs on the separate measures in Figure 3 cannot be used to assess the difference.

When seeing a figure with error bars, the first question to spring to mind should ask what the bars represent. Having

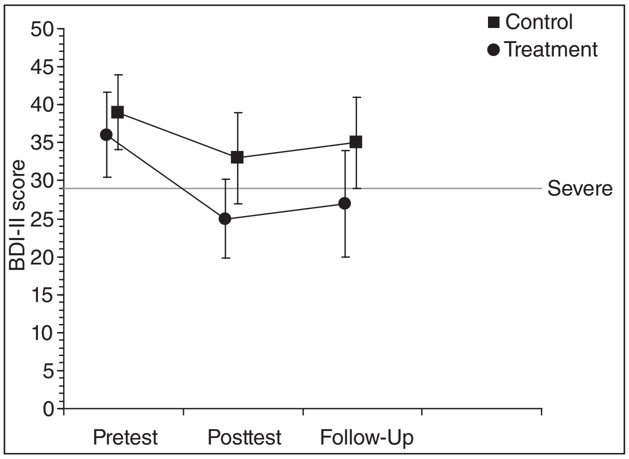

Figure 4 A fictitious randomised control trial (RCT). Beck Depression Inventory (BDI-II) pretest, posttest, and follow-up means are shown for the treatment and control groups, with 95% confidence intervals (CIs). Results for the two pretests are as in Figure 2, and for the treatment pretest and posttest as in Figure 3. The horizontal line marks a score of 29, the lowest score considered severe. Figure from the **Figure** page of Exploratory Software for Confidence Intervals (**ESCI**) **chapters 14–15**.

confirmed they are 95% CIs, the next question is about the design: independent groups or repeated measure? If independent groups, we can use the displayed CIs to assess the differences between the means, but if repeated measure, we need further information because there is no representation in the figure of the correlation between measures, and it is the size of this correlation that determines how sensitive the design is. Without knowing that, preferably by seeing the CI on the difference, we cannot interpret. Belia, Fidler, Williams, and Cumming (2005) reported evidence that many researchers do not understand the distinction, so if you do you can feel quietly superior. There is more about it in Cumming and Finch (2005).

RCT

Figure 4 shows the results for our example RCT, including follow-up scores. Which comparisons between means can we assess, with guidance of the CIs in the figure? The answer is any between-groups comparison, but *not* any within-group comparison. Therefore, we could interpret the difference between the two group pretest means by visualising the difference with its CI—as shown in Figure 2. The large overlap of the CIs tells us that we have no evidence of a difference, but the wide CI we visualise for the difference indicates that there is considerable uncertainty, and we must not conclude that the difference is necessarily zero or small. We *cannot*, however, use Figure 4 to assess the change from pretest to posttest in the treatment group because that is a within-group comparison, and we lack information about the CI on that difference—as displayed in Figure 3.

Figures like Figure 4 that combine between-groups and within-group independent variables (IVs) are common in psychology. The figure captions should always make clear the status of every IV pictured, and we need to take great care when inspecting such a figure to use the error bars to assess differences only where appropriate—meaning only between groups. To assess within-group comparisons, we need the CIs on the differences, but adding several floating difference axes to Figure 4 would quickly create a mess. One solution is a separate figure, and Figure 5 presents the (posttest—pretest) means for each group, with CIs. We can interpret those differences and CIs, and can also compare the two because they are independent—one for each group. We could visualise the difference of the differences, with its CI, or we could note that the two CIs are close to just touching, so there is reasonable evidence of a difference, with p being approximately .01. In fact $p = 0.008$, which agrees with the $p < 0.01$ stated in the NHST format report at the start of this article.

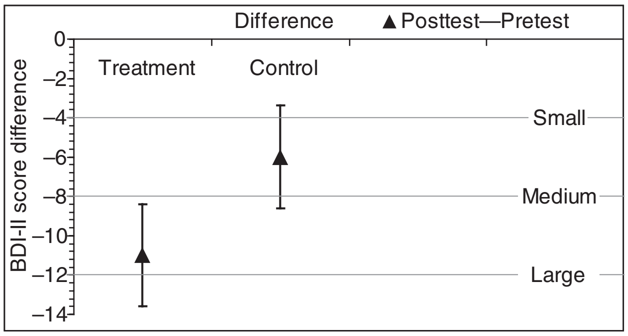

Figure 5 Mean posttest minus pretest differences for the randomised control trial (RCT) shown in Figure 4, with 95% confidence intervals (CIs). The horizontal lines mark changes from pretest to posttest considered clinically to be small, medium, and large. Figure from the **Figure** page of Exploratory Software for Confidence Intervals (**ESCI**) **chapters 14–15**.

In Figure 4, there are lines joining the repeated-measure means. This is a useful convention, worth using, and advocating. It is not, unfortunately, universally followed, so when seeing a figure in a journal we cannot rely on whether or not means are joined to indicate IV status.

RCT designs can be analysed in several ways, a common approach being an overall analysis of variance with focus on the groups-by-time interaction, perhaps followed by post hoc tests. However, Rosenthal and Rosnow (1985) and Rosenthal, Rosnow, and Rubin (2000) argued that such interactions are hard to interpret convincingly and that, in most cases, it is better to examine comparisons chosen to be most relevant to the researcher's aims. It is most convincing if a limited number of comparisons are nominated in advance of seeing the data. Fidler, Faulkner, and Cumming (2008) described how to take that approach to presenting and interpreting an RCT, using figures like Figures 4 and 5.

Interpretation of ESs and CIs

Step 5 requires us to interpret the ESs and CIs we report. This should include statements about size, importance, and the clinical or other practical implications, as appropriate for the context and research aims. We should go beyond saying simply that the true advantage of treatment over control is most likely between about 1 and 9 Beck points. It can be useful to support our interpretive statements by labelling in figures the reference ES values we consider large or small, or of particular degrees of clinical importance. Figure 4 marks a score of 29 as severe, because ranges of BDI-II scores are interpreted: scores of 0–13, 14–19, 20–28, and 29–63 are considered minimal, mild, moderate, and severe levels of depression, respectively (Beck et al., 1996). Figure 5 marks various sizes of change as small, medium, and large. Such reference labels can reflect the clinical judgement of the researcher, or may be values widely agreed within a research field or a clinical area. Different reference values may be appropriate for different populations or in different contexts. We need more such agreed reference values, and it will be a valuable development if more emerge as the new statistics become more widely used.

Noting the reference values, Figure 5 might prompt us to conclude that the control group showed a small-to-medium improvement, whereas the treatment group's improvement was medium-to-large, even approaching very large. Such interpretive statements might seem vague and subjective, especially compared with the apparently decisive "highly significant". Yes, they may be somewhat subjective, depending on how widely agreed the reference values are, but readers have the full information needed to make their own interpretations. A statement like "small-to-medium" captures not only the ES but also the extent of the CI, and thus reflects the extent of uncertainty in the data; no more precise statement can be justified. Interpretation is a matter of informed judgement in context, and researchers should trust their expertise when interpreting ESs and CIs.

Here is another example: "The reading age of 6 year olds increased by 4.5 months, 95% CI [1.9, 7.1], which is a large and educationally substantial increase. That of 4 year olds increased by only a negligible 0.6 months [–0.8, 2.0] . . ."

We are given the ESs (increases in reading age, in months) and CIs, and an educational interpretation. Even better would be to have the CI on the difference between the two independent groups (6 year olds and 4 year olds), although we can note that the two CIs are close to just touching (the lower limit for the older children, 1.9, is almost the same as the upper limit for the younger group, 2.0) and so we have evidence of a difference. A figure like Figure 2 would also be useful.

A Standardised ES: Cohen's *d*

Cohen's *d* is a number of SDs. To transform an original-units ES into *d*, we select an appropriate SD then divide. The general formula is

$$d = \frac{\text{ES in original units}}{SD}$$

where *SD* is our chosen standardiser. Standardised ES measures are valuable because they allow us to compare results obtained using different original-units measures.

Choosing the standardiser is a matter of judgement, and amounts to choosing the units in which *d* is expressed. Which *SD* makes most conceptual sense as a reference unit, in the context? If we know a suitable population *SD*, that is almost certainly the best choice. A well-normed intelligence quotient (IQ) test, for example, may scale its scores to have $SD = 15$ in some large reference population. If we judge that population to be a suitable referent for our research, we could regard 15 as a population value and our standardiser. Then, if we observed a difference of six IQ points, we could express that as $d = 6/15 = 0.40$.

However, we rarely know a suitable population *SD*, and must estimate from our data the population *SD* we want as our standardiser—usually that estimate will simply be some sample *SD* calculated for our data. Our value of *d* will then be a ratio between an original-units ES and that *SD* estimate. This makes interpretation difficult because two values of *d* may differ because the ESs differ, and/or the *SD* estimates differ. Even if we are using the same original-units measure, *SD* estimates will differ because of sampling variability, and so a *d*-value of 0.5 might reflect ES = 10 and $SD = 20$, or ES = 6 and $SD = 12$, or any number of other pairs of ES and *SD* values. Cohen's *d* is measured on a *rubber ruler*, in the sense that the measurement unit is subject to sampling variability—the ruler stretches in or out as our *SD* estimate varies over samples. We should always bear in mind this ambiguity of *d*, and should try to minimise the problem by finding as good an estimate of our chosen *SD* as we can.

For the two-independent-groups design, the pooled within-groups *SD*, as used to calculate independent-groups *t*, is usually the best choice as standardiser. For the results shown in Figure 2, ES = −3.0 and pooled $s = 14.0$, so we can express the difference between the two pretest means as $d = -3.0/14.0 = -0.21$. It is our arbitrary choice whether we use (treatment-control) or (control-treatment) as our ES, but we need to state which and be consistent. Here, I am using (treatment-control) as in Figure 2.

For the paired design, the *SD* of the pretest is usually the best choice. For the results shown in Figure 3, ES = −11.0 and the *SD* of the pretest is 15.0, so we can express the change from pretest to posttest for the treatment group as $d = -11.0/15.0 = -0.73$.

Considering our example RCT as a whole, the best standardiser is probably the pooled *SD* for the two groups of pretest scores, which is $s = 14.0$. By pooling over the two groups, we are likely to get a better estimate of the *SD* in the population from which we assume our treatment and control groups came. For consistency, we should use that *SD* estimate for any calculation of *d* for any comparisons in Figures 2, 3, or 4, whether relating to between-groups or within-group comparisons. Using this *SD* as standardiser for the difference shown in Figure 3 would give $d = -11.0/14.0 = -0.79$. This *d* differs from the value we calculated in the previous paragraph because we are now using a different standardiser.

In general, unless samples are large, our *SD* estimate is likely to have low precision as an estimate of the population *SD* we want as our standardiser. In other words, a repeat of the experiment is likely to give a different *SD* value, simply because of sampling variability—the rubber ruler will stretch in or out. We usually cannot avoid using the *SD* from our study as standardiser, but if there are several similar studies—ours or previously published—that all provide estimates of what we judge to be the same underlying population *SD*, then we should consider a pooled estimate of that *SD*, based on combining data from all the studies. That should give a much more precise *SD* estimate to use as standardiser—a thicker, less stretchy rubber ruler.

The population ES corresponding to Cohen's *d* is δ (delta). It is natural to use the *d* calculated from our data as our estimate of δ, but, unfortunately, *d* is a biased estimate of δ, and tends to overestimate δ, especially for small *N*. A simple correction is needed to our *d*, to obtain d_{unb}, which is the unbiased estimate of δ, and almost always what we should prefer. The formula was provided by Cumming, (2012b, pp. 294–295). When our standardiser is an *SD* estimate calculated for a sample with $N = 30$, *d* overestimates δ by about 3%. Therefore, the $d = -0.73$ we calculated earlier for the difference between pretest and posttest shown in Figure 3, using the pretest *SD* as standardiser (for which $N = 30$), would be reduced in absolute size by 3% to give $d_{\text{unb}} = -0.71$. For smaller *N*, the correction is greater. For *N* larger than around 50, the correction is 1.5% or less, so can usually be ignored.

A variety of terms are used for various forms of *d*. You may encounter Hedges' *g*, Glass' Δ (upper case delta), and, especially in medicine, standardised mean difference. The different terms reflect different choices of standardiser and whether or not the correction for bias has been made. The trouble is that these terms are used inconsistently, and the most common usage of some has changed over the years. The safest policy is to use the generic terms Cohen's *d*, and d_{unb} for the unbiased form, and to make very clear how it has been calculated—most notably, what standardiser was used? When seeing values of *d*, we need to know how they were calculated or we cannot interpret them.

Values of *d* are most commonly interpreted by using the reference values suggested by Cohen (1969): 0.2, 0.5, and 0.8 for small, medium, and large, respectively. Cohen emphasised, however, that interpretation should be based on informed judgement in the context, and these reference values used only if they are judged appropriate. In some cases, $d = 0.1$ may save lives; in others, $d = 1.0$ may be routine and only of moderate interest. As ever, researchers should use their informed judgement to interpret ESs.

If reporting *d*-values we should also report CIs. Because *d* is usually the ratio of two values, both calculated from data and thus both including sampling error, its distribution is complex, and there is no simple way to calculate accurate CIs. One solution is to use specialised software, such as the **CI for d** page of **ESCI chapters 10–13**. Use that page to find that the difference between the treatment and control pretests in our example, as shown in Figure 2, is $d = -0.21$ [−0.72, 0.30]. Note that the CI on *d* is usually slightly asymmetric, meaning the lower and upper arms are a little different in length. Another solution is to use the good approximation described by Cumming and Fidler (2009). There is more about all these aspects of Cohen's *d* in Cumming (2012b, Chapter 11).

Correlation

Pearson correlation, *r*, is an ES that measures the linear component of the relation between two variables, X and Y. It can take values from −1 (perfect negative correlation) through 0 (no correlation) to 1 (perfect positive correlation). It is very widely used in psychology, and most psychologists are probably familiar with scatterplots, which provide a useful picture of the relation between X and Y that is often summarised by the value of *r*.

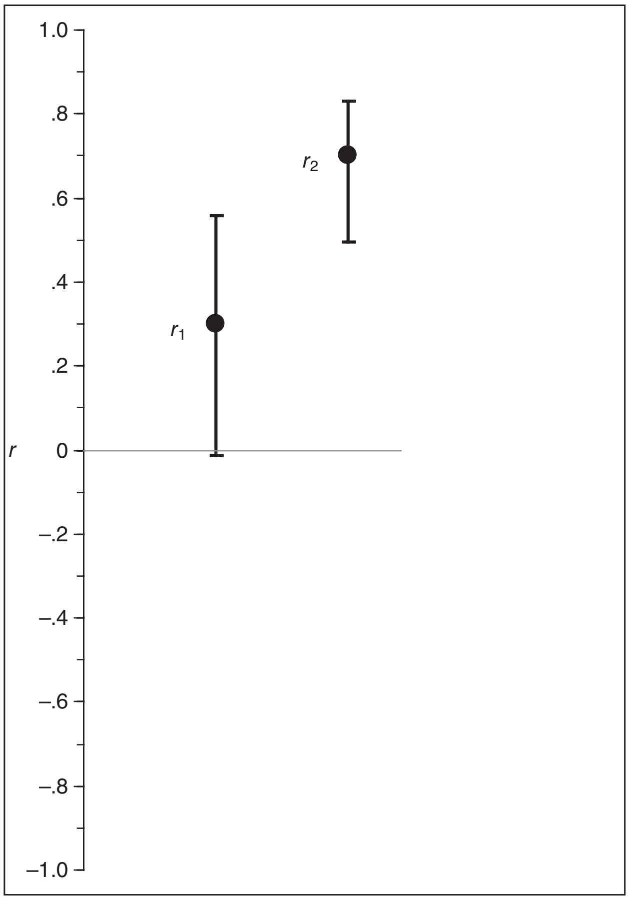

Figure 6 Two correlations, each with $N = 40$, and their 95% confidence intervals (CIs). The first is $r_1 = 0.30$ [–0.01, 0.56] and the second is $r_2 = 0.70$ [0.50, 0.83]. Figure from the **Two correlations** page of Exploratory Software for Confidence Intervals (**ESCI**) **chapters 14–15**.

To calculate a CI on *r* (or to use NHST), we need to assume X and Y have a bivariate normal distribution in the underlying population. Because *r* is bounded and must lie in the range (−1, 1), CIs on *r* are usually asymmetric. Figure 6 shows CIs on correlations of $r_1 = 0.30$ and $r_2 = 0.70$, each for groups with $N = 40$. For smaller *N*, CIs will of course be longer, and for larger *N* they will be shorter. The figure illustrates how, for a given *N*, CIs on *r* vary—when closer to 1 (or −1) the CI is shorter and more asymmetric. The CIs may seem surprisingly long, even with samples as large as 40; the CI for 0.30 even includes zero, so $r_1 = 0.30$ is not statistically significantly different from zero, at the 0.05 level, when $N = 40$. When reading any statement about a correlation being significant, it is worth bringing to mind the approximate CI, which is likely to be long or very long unless *N* is large.

There is a particular problem with NHST applied to *r*. Any time we see a *p*-value, we should ask what null hypothesised value was used to calculate *p*. For *r*, the most common choice is zero, and so a low or very low *p*-value may justify rejecting the

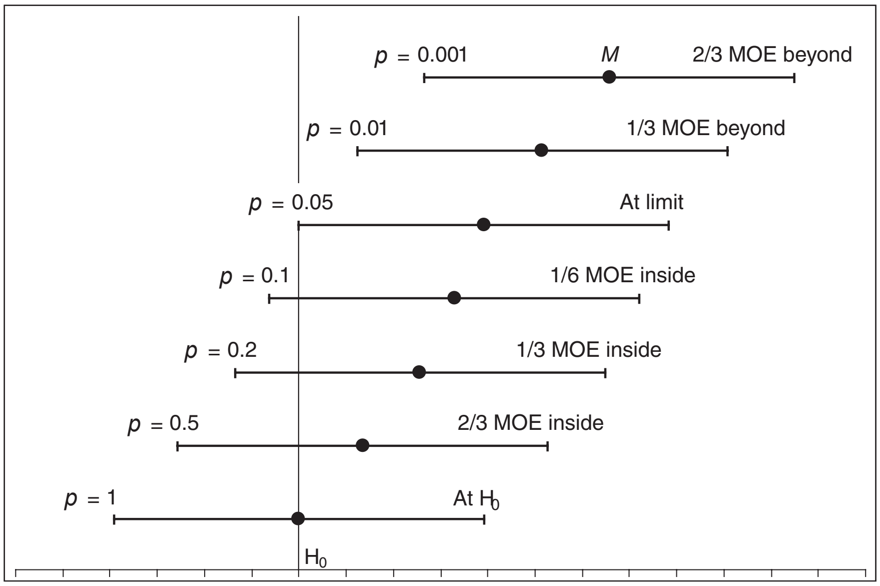

Figure 7 The relation between 95% confidence intervals (CIs) and the two-tailed *p*-value. The *p*-values at left correspond to the positioning of the CI in relation to the null hypothesised value indicated by the vertical line. The labels at right state approximately where the H_0 value lies in relation to the CI, in terms of margin of error (MOE), which is the length of one arm of the CI.

null hypothesis of zero. In many situations, however, a correlation of zero is a terrible choice of null hypothesis. When *r* is used, for example, to measure reliability or validity, a value of 0.7, 0.8, or even 0.9 might be considered routine, or even disappointing. Claiming that a value of, say, 0.7 is very highly significant, just because the null hypothesis of zero can be rejected with a tiny *p*-value, is irrelevant and perhaps misleading. If $N = 40$, Figure 6 shows the CI is [0.50, 0.83], and this, as usual, is much more informative than any statement about significance. Assistance in calculating CIs on *r* and preparing figures is provided by the **Correlations** and **Diff correlations** pages of **ESCI Effect sizes**.

Research Reported Using NHST

Even when using the new statistics ourselves, we need to be able to understand published research reported using NHST. Fortunately, with a little practice it is easy in many cases to visualise the CI for a result reported with a *p*-value. Figure 7 illustrates the translations we need. The third CI from the top illustrates the well-known relation that, if a 95% CI lands with either end equal to the H_0 value, the two-tailed *p* is .05. As the CI moves further from H_0 (the two top CIs in Figure 7), it offers stronger evidence against H_0 and *p* gets smaller. Moving down in Figure 7, the CI moves so that H_0 is closer to *M*, at the centre of the CI, and *p* increases.

The labels at the right in Figure 7 give approximate benchmarks that are worth remembering. They can be used in either direction: to read a *p*-value from a CI, or to visualise the CI, given *p*. However, Coulson, Healey, Fidler, and Cumming (2010; tiny.cc/cisbetter) reported evidence that, at least in some common situations, researchers who see results presented as CIs are much more likely to interpret the results correctly if they think in terms of estimation than if they invoke NHST. Therefore, eyeballing a *p*-value is far from the best way to interpret a CI, but it is an option, and Figure 7 gives a range of examples.

I include Figure 7, however, to provide an estimation way to interpret results presented using NHST. We need to know the H_0 value (perhaps zero), the mean, and an exact *p*-value, then we can visualise the extent of the CI and where it falls in relation to H_0. Consider, for example, our initial NHST example: We know the ES is 5 (the difference between 11 and 6, the mean improvement scores for the two groups) and assume $p = .01$. (The *APA Manual* advises that, if using NHST, exact *p*-values should be reported, rather than only relative values such as $p < .01$. Having the exact *p*-value would allow us to visualise the CI more accurately.) The *p*-value of .01 directs us to the CI second from top in Figure 7, which we can remember as positioning H_0 about one third of MOE beyond the CI. We have H_0 of zero and $M = 5$, so MOE is a little less than 4, say 3.7. Our visualised CI is thus about [1.3, 8.7], and we can interpret the result by interpreting that interval, which gives us a much more realistic idea of the uncertainty in the result than the statement about a highly significant difference. Note that our visualised CI is quite similar to the accurate CI reported in the estimation format.

Suppose ES = −3.0 and $p = .42$. Can you visualise the CI? The closest reference CI is second bottom in Figure 7, for $p = .50$. Our *p* is a little smaller, so H_0 is positioned a little less than about two thirds of MOE in from a CI limit. Noting that −3 is the distance from H_0 to *M*, we can estimate MOE to be about 7, and therefore our visualised CI is about [−10, 4]. The ES and *p*-values are actually those for the difference shown in Figure 2, and our visualised CI is quite similar to the accurate CI shown on the difference in that figure. It is worth remembering the patterns in Figure 7, and practising visualising CIs given only an

ES and a p-value—this may often be the most insightful way to interpret results reported using NHST.

CIs Are Often Disappointingly Long

The observation that CIs are often disappointingly long should prompt three responses. First, we must not let that count against CIs because they are accurately reporting uncertainty. Do not shoot the messenger! Second, we should make every effort to design better, more sensitive experiments, using every strategy recommended in books on experimental design. For example, use matching or repeated measures where possible, use larger samples, choose more reliable measures, and improve experimental control. In this way, the CIs in future experiments should be shorter, and therefore more informative.

Third, recognise that rarely does a single experiment answer a research question adequately. Almost always, we should be thinking in terms of meta-analysis. Where possible, we should use meta-analysis to combine our study with previously published similar studies. In any case, we should report our study sufficiently completely to assist its inclusion in future meta-analyses. This may prompt us, for example, to report our results using r and/or d as ES measures, probably in addition to reporting ESs in original units. These two ES measures, r and d, are the most commonly chosen in psychology for carrying out meta-analysis.

Practical Challenges in Using the New Statistics

We need to be realistic, and report our results so that journal referees and editors will publish our work. Many still expect NHST, even if many journals now state that ESs must always be reported. Here are some suggestions:

APA Manual

The *Manual* (APA, 2010) makes unequivocal statements calling for use of the new statistics. It specifies a format for reporting CIs, and includes numerous CI examples, for a wide range of measures and situations. Cumming et al. (2012) discussed the *Manual*'s recommendations further. Strong support by the *Manual* is a vital legitimation of statistics reform, and encouragement for researchers to use the new statistics.

Explain and Justify

In a postgraduate thesis, it may be worth a paragraph or more to explain the new statistics data analysis and reporting strategy to be used throughout. In a journal manuscript, a sentence or two may have to suffice. In the thesis, perhaps describe the six-step estimation strategy, mention the *APA Manual*, and cite references in support. Kline (2004, Chapter 3) can assist. Any examiner or journal referee should respect a carefully justified data analysis strategy, even if different from their expectation.

Evidence should Determine

In any area of science, the data should ultimately determine. Choice of statistical practice should likewise be guided by evidence, for example evidence that NHST is poorly understood (Haller & Krauss, 2002; tiny.cc/nhstohdear) and often misused, and that estimation can often be better (Coulson et al., 2010; tiny.cc/cisbetter).

More Complex Situations

This article considers only means and two other ES measures, and simple designs. What about more complex designs, multivariate research, model fitting? These are good questions, but the flaws of NHST are just as damaging in complex situations, and the prospects of finding new statistics alternatives are improving. For example, Tabachnick and Fidell (2007) is a widely used textbook that includes advice on calculating and using CIs in a wide variety of multivariate situations.

If an Editor Insists

If an editor insists, I may choose to include some p-values, but I will make them as inconspicuous as possible and will not refer to them when discussing and interpreting my results. I will not remove essential new statistics reporting.

Be Strong

Adoption of the new statistics, or other techniques better than NHST, is not a matter of fashion or mere personal preference. Numerous scholars have published cogent critiques of NHST and explanations of the damage it does, and almost no defences of it have been published. It is clear that best practice, which should be the goal of every scientist, is almost always to avoid NHST and to use better techniques instead. Let the arguments of Kline (2004, Chapter 3) strengthen your resolve. Perhaps scan quotes from scholars who have written about statistical reform: See tiny.cc/nhstquotes. I will finish with an example quote more than half a century old, but just as appropriate today: "The traditional null-hypothesis significance-test method . . . is here vigorously excoriated for its inappropriateness" (Rozeboom, 1960, p. 428).

References

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

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of beck depression inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, *67*, 588–597. doi:10.1207/s15327752jpa6703_13

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396. doi:10.1037/1082-989X.10.4.389

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement*, *1*, 26. doi:10.3389/fpsyg.2010.00026 Retrieved from tiny.cc/cisbetter

Cumming, G. (2012a). The new statistics: What we need for evidence-based practice. *InPsych*, *34*(3), 20–21. Retrieved from tiny.cc/tnsinpsych

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

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 15–26. doi:10.1027/0044-3409.217.1.15

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*, 138–146. doi:10.1111/j.1742-9536.2011.00037.x

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology*, *177*, 7–11. doi:10.1083/jcb.200611141 Retrieved from tiny.cc/errorbars101

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist*, *60*, 170–180. doi:10.1037/0003-066X.60.2.170 Retrieved from tiny.cc/inferencebyeye

Fidler, F., Faulkner, S., & Cumming, G. (2008). Analyzing and presenting outcomes: Focus on effect size estimates and confidence intervals. In A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 315–334). New York: OUP.

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

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(1), 1–17. Retrieved from tiny.cc/nhstohdear

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: APA Books. Chapter 3. Retrieved from tiny.cc/klinechap3

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research. A correlational approach*. Cambridge, UK: Cambridge University Press.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. doi:10.1037/h0042040

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.