Confidence Intervals

Better Answers to Better Questions

Geoff Cumming and Fiona Fidler

La Trobe University, Melbourne, Victoria, Australia

Abstract. Most questions across science call for quantitative answers, ideally, a single best estimate plus information about the precision of that estimate. A confidence interval (CI) expresses both efficiently. Early experimental psychologists sought quantitative answers, but for the last half century psychology has been dominated by the nonquantitative, dichotomous thinking of null hypothesis significance testing (NHST). The authors argue that psychology should rejoin mainstream science by asking better questions – those that demand quantitative answers – and using CIs to answer them. They explain CIs and a range of ways to think about them and use them to interpret data, especially by considering CIs as prediction intervals, which provide information about replication. They explain how to calculate CIs on means, proportions, correlations, and standardized effect sizes, and illustrate symmetric and asymmetric CIs. They also argue that information provided by CIs is more useful than that provided by p values, or by values of Killeen's p_{rep}, the probability of replication.

Keywords: confidence intervals, effect sizes, estimation, p_{rep} probability of replication, p values

> How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between *A* and *B* and the error associated with our estimate; knowing that *A* is greater than *B* is not enough. (Kirk, 1996, p. 754).

> Psychology will be a much better science when we change the way we analyze data. (Loftus, 1996, p. 161)

What is the lifetime of a muon? What pollution level is tolerated by this species of fish? What is the melting point of this new material? These are typical questions across science, and the most useful answer in each case is quantitative. The materials scientist expects an answer to the third question to be something like 155.6 ± 0.4 °C, where the 155.6 is the best point estimate of the melting point, and the 0.4 is an error margin that expresses the precision of the point estimate. Similarly, in psychology the most useful answer to a very wide range of research questions is quantitative, and comprises a point estimate and information about the precision of that estimate. In the early decades of scientific psychology, seeking such quantitative information was the natural thing for researchers to do, and the answers filled the early journals. However, for the last half century or so, data analysis in psychology has been dominated by null hypothesis significance testing (NHST), and remains so currently (Cumming et al., 2007). NHST focuses attention not on estimating amounts, but on the impoverished question of whether or not some effect of research interest is zero.

The severe deficiencies of NHST and its damaging effects on research decision-making have been widely documented, notably by Kline (2004, chapter 3). Kline identified 13 widespread but seriously wrong beliefs about p values and the way they are used by psychologists conducting NHST. We will not reiterate the catalog of NHST shortcomings here, but will focus on one, which may be its most insidious: NHST encourages dichotomous thinking (Kline, p. 76). Examining a p value leads the research psychologist to reject the null hypothesis of no difference – if, for example $p < .05$ – or not to reject the null. Consequently, researchers come to formulate theories in terms of whether or not certain variables do or do not have some influence on other variables. At most there may be prediction of the direction of influence, but nothing is said about the strength of relationships; there is no quantitative modeling. What a blinkered, shallow view of science! NHST is the culprit, and it has hobbled the theoretical thinking of psychologists for half a century.

Our hope is that a major shift by psychologists from NHST to estimation, meaning the reporting of point estimates with information about precision, will not only give better, more useful quantitative answers to questions, but also will lead psychologists to ask better questions. Confidence intervals (CIs) are interval estimates that conveniently combine information about point estimates and their precision. We hope that reporting CIs wherever possible, and using them as the basis for interpretation of data, will lead to a theoretically richer, and a more quantitative and precise discipline, as called for by Meehl (1978), and Loftus (1996) in the quotation at the start of this article. Later we will, in addition, find fault with p values because of their unreliability: A repeat of an experiment is likely to give a p value very different from that given by the original.

Rosnow and Rosenthal (2009) advocate the reporting and interpretation of effect sizes (ES) and the use of CIs, mainly in the context of some ongoing use of NHST. We

DOI 10.1027/0044-3409.217.1.15

support their argument, but in this article we take a complementary approach by advocating use of ESs and CIs without necessarily any reference to NHST. We start by discussing ESs, then the interpretation of simple CIs. We compare the information conveyed by a CI with that conveyed by a *p* value, and also a p_{rep} value (Killeen, 2005), and argue that the CI gives better information on which to base interpretation of results. We then discuss and illustrate CIs on a number of types of measure, including simple means, proportions, correlations, and standardized ESs, and refer to some resources to assist in calculating CIs. Our aim is to encourage and support the highly desirable transition by psychology from the dichotomous thinking of NHST to the CI world in which psychologists ask better, more quantitative questions, and CIs provide better, more informative answers.

Effect Sizes

An effect size (ES) is simply the amount of anything of interest. It can be as familiar as a mean, a difference between means, a median, a percentage increase, a correlation, a frequency, a regression slope, or a proportion of variance. It may be expressed in original measurement units – the units in which the dependent variable was originally measured (e.g., ms, kg/month, or cm^2) – or in some standardized units (e.g., ω^2, β, Cohen's *d*), or it may be a units-free measure (e.g., *r*, b, proportion, odds ratio). We can consider a population ES, and also the sample ES that we use to estimate it.

The *Publication Manual* of the American Psychological Association (APA) stated that "it is almost always necessary to include some index of ES or strength of relationship . . ." (APA, 2001, p. 25). Many journals now state in the instructions to authors that ESs should be reported for all effects of interest. That these statements need to be made at all is an indicator of the extent to which NHST has captured psychology. The materials scientist would be highly puzzled by any stated journal requirement that she must report the value of the melting point: Of course! Measuring that was the object of her careful bench work! Similarly, before the middle of the 20th century and the rise of NHST, no psychology journal needed to state that authors must report the means, or other quantitative results of their experiments – everyone understood that the primary purpose of an empirical article was to report such values.

Only under the oppressive influence of NHST could a finding be reported merely as statistically significant ($p < .05$), or as *ns*, with no mention of means or other descriptive statistics. Such reporting is totally inadequate for the reader trying to understand the results, the researcher considering a replication or, most damagingly, the researcher seeking to include the results in a future meta-analysis. Of course the ES value must be reported for every effect studied, whether or not any particular *p* value is obtained, and whether or not statistical significance can be claimed.

Further, CIs should be reported for ES values wherever possible. The APA Task Force on Statistical Inference (Wilkinson & the Task Force on Statistical Inference, 1999) stated that: "Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients . . ." (p. 599); and "In all figures, include graphical representations of interval estimates whenever possible" (p. 601). We turn now to CIs, starting with a simple example.

Confidence Intervals

CIs: A Simple Example

Suppose we are investigating the effectiveness of a workshop designed to increase well-being. Sixty-four participants have been randomized into two groups, each with $n = 32$, one group to take the workshop, and the other to serve as a waiting-list control. We use a (fictitious) well-standardized test of well-being, and assume the population *SD* is $\sigma = 20$ for both groups and that population scores are normally distributed. We obtain well-being scores for our experimental group after taking the workshop, and for the control group, and find M_{diff}, the difference between the two means, to be 11.08. This value is plotted as the gray dot near the bottom of Figure 1, just above the distribution, and the 95% CI is shown on this mean. The *margin of error* (MOE) of this CI is the length of either arm of the CI, or half the total width of the interval. It is calculated as MOE = $z_{.95} \times \text{SE}_{\text{d}}$, where $z_{.95}$ is the critical value of *z* required for a 95% CI, and SE_{d} is the standard error of the difference between the two group means. Now $\text{SE}_{\text{d}} = \sigma\sqrt{}\text{v}(1/n + 1/n) = 20/4 = 5$, and so MOE = 1.96 × 5 = 9.80. The result could be reported in text as "the difference in mean well-being scores between the two groups was 11.08, 95%CI [1.28, 20.88]." The two values in brackets are the *lower limit* of the 95% CI, calculated by subtracting the MOE (9.80) from the point estimate (11.08), and the *upper limit*, which is the point estimate plus the MOE.

Figure 1 shows M_{diff} values and their CIs for a further 24 replications of our experiment, the replications being identical but using independent random samples of participants. The computer simulation used to generate Figure 1 assumes our workshop increases well-being by 10 points, or 0.5 σ, meaning there is a true population difference of half a *SD* between the two populations. The dotted vertical line labeled μ marks this true population difference, and μ is the population parameter we are estimating. In terms of Cohen's (1988) somewhat arbitrary but useful ES reference values (0.2, 0.5, and 0.8 *SD* for small, medium, and large effects, respectively), this is a medium-sized effect. Figure 1 illustrates how the CIs jump around on repeated sampling, approximately symmetrically either side of the pop-

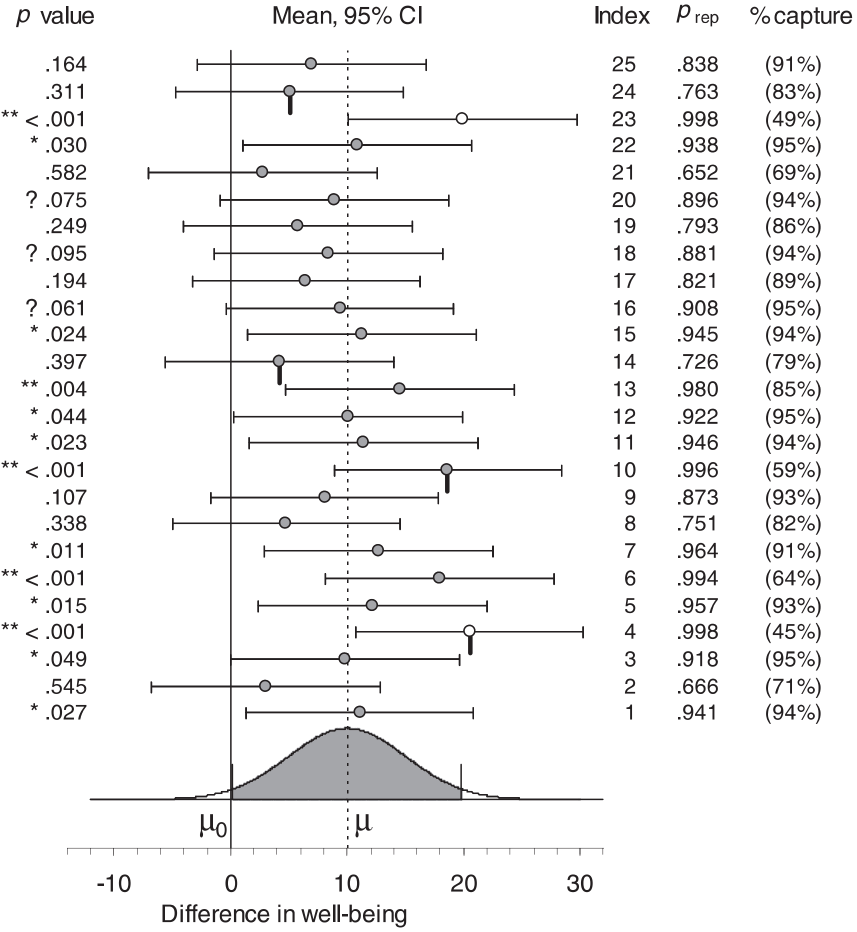

Figure 1. Means and 95% confidence intervals (CIs) for 25 simulated replications of a two independent group experiment, with $n = 32$ in each group. The means are values of M_{diff}, the difference between the mean well-being scores for the two groups. The underlying populations are assumed to be normally distributed, each with known *SD* of $\sigma = 20$, z being used to calculate the CIs. The dotted vertical line marks $\mu = 10$, which is assumed to be the difference between the population means, so the population effect size is 0.5s, a medium-sized effect. The curve near the bottom is the sampling distribution of M_{diff} values, and the shaded area between the two short verticals is the symmetric interval within which – in the long run – 95% of M_{diff} values will fall. The column at left shows for each experiment the two-tailed p value for a z test of the null hypothesis, $\mu_0 = 0$, of no difference between the two population means. The value of μ_0 is marked by the solid vertical line. In the long run, 95% of CIs are expected to capture μ, and here all but numbers 4 and 23, which are marked by open circles, do so. The percentage of future replication means that each CI will capture is shown in parentheses in the rightmost column; this is the percentage of the area under the lower curve that is encompassed by that CI. The short thick vertical lines on means 4, 10, 14, and 24 indicate those means that fell outside the previous CI; 95% CIs are expected to capture on average about 83% of replication means. The column second from the right shows for each experiment the value of Killeen's p_{rep}, the probability a replication will give a same-sign M_{diff}. Note the very large variation over the replications in both p values and p_{rep} values.

ulation parameter μ. Any textbook that explains CIs correctly includes a figure that illustrates this jumping around, and explains that the CI, commonly and in this case 95%, is the long-run percentage of intervals that fall so they include the true population value. In Figure 1, two of the intervals, numbers 4 and 23 (see the Index column), do not capture μ; they are marked by open circles. When examining Figure 1, keep in mind that it is a simulation and in real life the researcher does not know μ and has only a single sample. The 25 replications illustrated are a sample from an infinite sequence of potential results, of which 95% give CIs that include μ, and 5% CIs that miss. In real life, unfortunately, the small number of CIs that do not include the population value cannot be identified by open circles!

Thinking and Talking About CIs

Figure 1 illustrates the basic and correct way to think about a 95% CI, in terms of the long-run proportion of intervals that include μ. Any single interval either does or does not include μ, so the probability of capture is 1 or 0 and it is misleading to say "there is a .95 chance our 95% CI includes μ." We might say "we are 95% confident our interval includes μ," and can bear in mind that in a lifetime of research, and reporting of numerous 95% CIs in a wide range of situations, about 95% of our reported intervals will capture the population parameter being estimated.

The normal distribution near the bottom of Figure 1 is the sampling distribution of M_{diff}; it shows that M_{diff} most often falls near μ, and progressively less often falls at greater distances from μ. The shaded area extends from 1.96 × SE_d = 9.80 below μ to the same distance above μ; 95% of M_{diff} values will fall in that interval. The ordinate (i.e., height) of the curve at μ is about 7 times the ordinate at either end of the shaded area, and so M_{diff} is about 7 times as likely to fall in a tiny interval at μ as in a tiny interval of the same length at the lower (or upper) end of the shaded area. In other words, the relative likelihood of the estimation error $|M_{\text{diff}} - \mu|$ being zero is about 7 times greater than it being 1.96 × SE_d, which is the maximum estimation error for means in the central 95% of the sampling distribution. In simple terms, M_{diff} is most likely to land near μ, and the curve in Figure 1 illustrates how the relative likelihood of M_{diff} values drop further from μ. Consistent with this expectation, the intervals shown in Figure 1 fall so that most often the sample mean is fairly close to μ; intervals less often fall so that their lower or upper limit is close to μ.

Now focus on any single 95% CI by itself and consider that, as in real life, we don't know μ. Small estimation errors are more likely than large, so values close to our M_{diff} are our best bet for μ, and values near either limit of our CI are, by a factor of about 7, less good bets. In fact, the curve in Figure 1, if centered on our M_{diff}, rather than on μ as in Figure 1, gives the relative likelihood of the various values in and beyond our CI being the true value of μ. Values close to the point estimate M_{diff} are the most plausible for μ. Values inside our interval but out toward either limit are progressively less plausible for μ. Values just outside the interval are relatively implausible, but not impossible as values for the parameter. The curve in Figure 1 is simply a normal distribution with standard deviation equal to SE_d. The graphic on the right in Figure 2 is that distribution (and its mirror image) centered on the sample mean. Figure 2, thus, illustrates how relative likelihood, or plausibility, varies across a 95% CI by comparing the conventional graphic and the bulging graphic on the right whose width is proportional to the likelihood that μ takes a particular value on the vertical axis.

The wording "the interval falls so that it includes μ" may seem awkward, but is intended to emphasize that it is the interval that is the variable, in contrast to the population parameter, which is an unknown fixed value. Saying "the probability is .95 that μ lies in this interval" is not recommended because it too easily gives the incorrect impression that μ is the variable.

A CI can, thus, serve as the basis for interpretation of results. Thinking of a CI in terms of various levels of plausibility is probably consistent with what the materials scientist would say if we quizzed her about how she interprets her result of 155.6 ± 0.4 °C: The true value is most likely quite close to 155.6, is less likely to be above or below this value by as much as 0.4 °C, and is quite unlikely to be further from this value. The variation of plausibility across a CI was illustrated and discussed further by Cumming (2007). Cumming and Finch (2005) discussed additional ways to think about CIs.

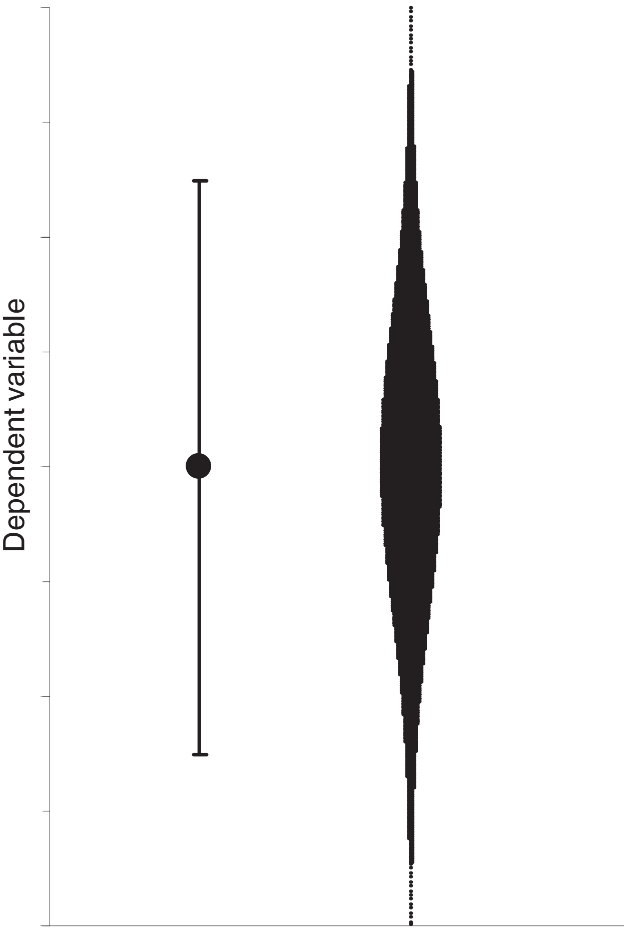

Figure 2. Two representations of a 95% CI. The conventional bars are on the left. The width of the graphic on the right is proportional to the likelihood of a value on the vertical axis being the true value of μ, the population parameter being estimated. The bulging figure thus indicates the relative likelihood that μ falls at any given value.

CIs as Prediction Intervals

The CIs in Figure 1 do bounce around, and a few are extreme, but the width of any interval does give an indication of the extent of bouncing around – and of where other means will fall. Knowing just one interval gives some idea of the whole infinite set of possible CIs, just 25 of which are illustrated in Figure 1. In other words, a CI can be regarded as a *prediction interval* for future means: A certain percentage of means of replication experiments will fall within that CI, and the percentage will be different for different CIs, depending on where a particular interval falls in relation to μ. The curve near the bottom of Figure 1 shows where replication means will fall. The area under this curve within the limits of a particular CI defines the proportion of replication means that will fall within that CI. The column at right gives the percentage of replication means that each CI will capture. Intervals whose M_{diff} falls close to μ (e.g., Experiments 3, 12, and 22) will capture close to 95% of future means because their extent is a near match to the shaded extent under the sampling distribution, which includes 95% of means. As sample means (M_{diff} values) fall progressively further from μ, the percentage of replication means captured drops because the corresponding areas under the sampling distribution progressively decreases. Experiments 15, 25, 13, and 2, in that order, illustrate the progression. Intervals that do not include μ lie entirely in either one or the other half of the sampling distribution, and so will capture less than 50% of future means: See Experiments 23 and 4.

The capture percentages in Figure 1 are shown in parentheses to indicate that knowledge of μ is needed for their calculation: In real life, in which μ is unknown, we cannot calculate the capture percentage for the single CI based on our data. Cumming, Williams, and Fidler (2004) and Cumming and Maillardet (2006) discussed further the capture of replication means by CIs, and provided illustrations and simulations to assist understanding. They showed that, for a normal population when σ is known, the average capture percentage is 83.4%, so, on average, a 95% CI will include the mean for about five out of six replications. They also showed that the average capture percentage is close to 83% when σ is not known, at least when *n* is 10 or more, and for a number of nonnormal population distributions. In each case the distribution of capture percentages is, however, strongly negatively skewed, as the values in the rightmost column in Figure 1 illustrate. For the normal, σ-known case the median capture percentage is 89.6%, so most 95% CIs capture around 85 to 95% of future means – including all of the Experiments 15–20 in Figure 1. A few happen to fall a little distance from μ and so will capture lower percentages, for example Experiments 2, 6, 10 and 14. Just 5% of intervals will not include μ and, therefore, will capture fewer than 50% of replication means – Experiments 4 and 23. Of course, we never know the capture percentage for our particular CI, unless we can discover μ or repeat the experiment many times.

In Figure 1, the short, fat vertical lines on some sample means show cases in which the previous CI did not include this M_{diff} value: See Experiments 4, 10, 14, and 24. In the long run, we expect 16.6% (i.e., 100% – 83.4%) of means to be labeled in this way. Cumming et al. (2004) reported evidence that many researchers in psychology, behavioral neuroscience, and medicine appear to believe that a 95% CI is likely to capture about 95% of future replication means. The previous paragraph, and the patterns of Figure 1, explain why such a belief would be mistaken: 95% of intervals will indeed capture μ, but only an interval that happens to fall so that M_{diff} is at μ will capture as many as 95% of replication means. All other intervals will capture fewer. Replication is at the heart of science, and CIs give information about replication because a CI can be thought of as a prediction interval: On average, a 95% CI will capture about 83% of replication means.

CIs and *p* Values

The solid vertical line in Figure 1 is labeled μ_0 and marks zero, the null hypothesis for a statistical significance test of the difference between groups. The two-tailed *p* values for a *z* test of the M_{diff} values against this null hypothesis are shown at the left. There is a simple relation between how a CI falls in relation to the μ_0 line, and the *p* value. If M_{diff} falls at zero, $p = 1.0$, and *p* drops as M_{diff} is progressively further from zero in either direction. In Figure 1, Experiments 2 and 21 have M_{diff} values closest to zero and, correspondingly, the largest *p* values. If one of the limits of a CI falls exactly at zero, $p = .05$, as very nearly happens for Experiment 3. If the interval includes zero, $p >.05$, and if the interval does not include zero, $p < .05$; Experiments 12 and 16 show *p* values just below and just above .05 for intervals whose lower limits are, respectively, just above and just below zero. Intervals distant from zero give small *p* values, most notably for Experiments 4, 6, 10, and 23, which all give $p < .001$. Cumming (2007) suggested benchmarks that can be used to estimate *p* by noting in a figure where a 95% CI falls in relation to a null hypothesized value.

The most striking feature of the column of *p* values in Figure 1 is their astonishing variability: from greater than .50 to less than .001. It seems that practically any *p* value might be given by a replication of our experiment. Repeat your experiment and you are likely to get an entirely different *p*. How, then, can the *p* value from a single experiment be a reliable basis for interpretation? That is an extremely good question!

Perhaps the experiment in Figure 1 is uncharacteristic in some way, and the great variability in *p* is an aberration? In fact, the statistical power of the experiment is .52, meaning this is the probability that $p < .05$, given sample size $n = 32$ for each group, and an ES of 0.5σ. Studies of statistical power over several decades (Maxwell, 2004, 148) have shown that, for a number of fields in psychology, published research has median power to detect a medium-sized effect of around .5. So our example is typical of many experiments published in psychology, and the variability of *p* illustrates the *p*-value lottery that determines the outcome of such an experiment. Even if you are lucky and obtain a small *p*, a repeat of the experiment is likely to give a very different value of *p*. Cumming (2008) discussed further the variability of p values, and how this extreme variability – or unreliability – makes it unwise to base research decision-making on p.

Note in Figure 1 that a CI can be calculated simply from our sample data; no null hypothesized value is required. By contrast, a p value also requires the specification of a null hypothesis – usually, as here, a value of zero – and p is a measure of our data in relation to that hypothesis.

Considering a single experiment, what information would you like to see published? Would you be satisfied to be told the p value, and perhaps M_{diff}? Would p and M_{diff} give a reasonable indication of the whole set of possible results, 25 of which are shown in Figure 1? More specifically, is a single p value a reasonable exemplar of the whole set, a reasonable representative of the infinite set of possible p values? Does it give information about what a replication is likely to give? We suggest that the extreme variation in p means that the answer to all these questions should be no. By contrast, does any single CI give some idea of the whole set? Certainly the intervals bounce around, but we suggest that any one of them gives a reasonable idea of the extent of bouncing, and so does give a notion of the whole set. Also, as we have seen, a CI gives information about replication. We conclude that, considering replication and the infinite set of possible results, the comparison of CIs and p values in Figure 1 provides a strong reason for preferring CIs. A 95% CI gives fuller and more comprehensible information on which to base conclusions than does a p value.

CIs and p_{rep}, Killeen's Probability of Replication

Killeen (2005) proposed that p values be replaced as the basis for inference by the *probability of replication*, which he labeled p_{rep}. He defined p_{rep} as the probability that a replication result is of the same sign as the original result, and presented an ingenious way to calculate p_{rep}, which does not require knowledge of the true population ES. In effect, p_{rep} is a weighted average probability of obtaining a same-sign replication result, where the average is taken over all possible population ESs, weighted by the likelihood a particular ES gave the original result. The proposal has generated considerable interest, criticism, and discussion, and in follow-up articles Killeen (2006, 2008) responded, and extended his proposal for p_{rep} to provide the basis for inference and scientific decision-making. Cumming (2005) argued that p_{rep} may be hard to understand correctly, and provided explanations and a simulation intended to assist. Killeen (2005) and Cumming also explained that, at least in the case of the simple z test, p_{rep} can be calculated as a transformation of the p value. High p_{rep} values correspond to small p values, $p_{\text{rep}} = .92$ corresponds to two-tailed $p = .05$, and $p_{\text{rep}} = .97$ corresponds to two-tailed $p = .01$.

Since July 2005 *Psychological Science* has encouraged authors to use p_{rep} rather than NHST as the basis for inference, and since then p_{rep} has become common in that journal. However, most authors seem to continue to use NHST, but with p_{rep} used simply as a surrogate for p, rather than as the basis for an alternative approach to inference based on thinking about replication, as Killeen advocated. We believe p_{rep} deserves further investigation, especially in the context of Bayesian approaches to inference. Figure 1, however, illustrates one additional property of p_{rep} – the way it varies over replication. Values of p_{rep} calculated for each experiment are shown in a column to the right. As we expect, low p goes with high p_{rep}, p close to .05 (e.g., Experiment 3) gives p_{rep} close to .92, and p close to .01 (e.g., Experiment 7) gives p_{rep} close to .97. The correspondence between the two measures means that p_{rep}, as well as p, shows very large variability over repeats of the experiment. Repeat your study, and both p and p_{rep} are likely to be very different from those given by the original study, and so neither of these measures seems promising as a basis for statistical inference. In other words, again, it seems much more informative to be given the CI for a single study then either the single p value, or single value of p_{rep}.

In our example we assumed σ known, to limit the complexity of Figure 1 and the comparisons across CIs, p, and p_{rep}. If σ is not known, s the sample *SD* is used as an estimate of σ, and t rather than z is used to calculate CIs and p values. CI width varies from experiment to experiment. However, CIs still provide information about replication (Cumming & Maillardet, 2006), p and p_{rep} values still show very large variability, and our comparisons and the conclusions in favor of CIs remain the same.

Calculation and Presentation of CIs

CIs for Means

The Appendix includes the basic formulas for calculating CIs for means, and for the other ES measures discussed in the following subsections. Altman, Machin, Bryant, and Gardner (2000) provided formulas and advice for a wider range of ES measures, especially those commonly used in medicine, such as odds ratios; they also provided software for carrying out many of the calculations.

Figure 3 illustrates means and 95% CIs for a two-way design with one repeated measure, a common design in psychology. The CIs in Figure 3, like any CI on any measure of an ES, can be interpreted in any of the ways mentioned above, or in other ways, including those discussed by Cumming and Finch (2005). For example, we can think of any mean and its CI in Figure 3 as one randomly selected from an infinite set, 95% of which will include the corresponding population mean; or as a prediction interval that will, on average, include about 83% of replication means;

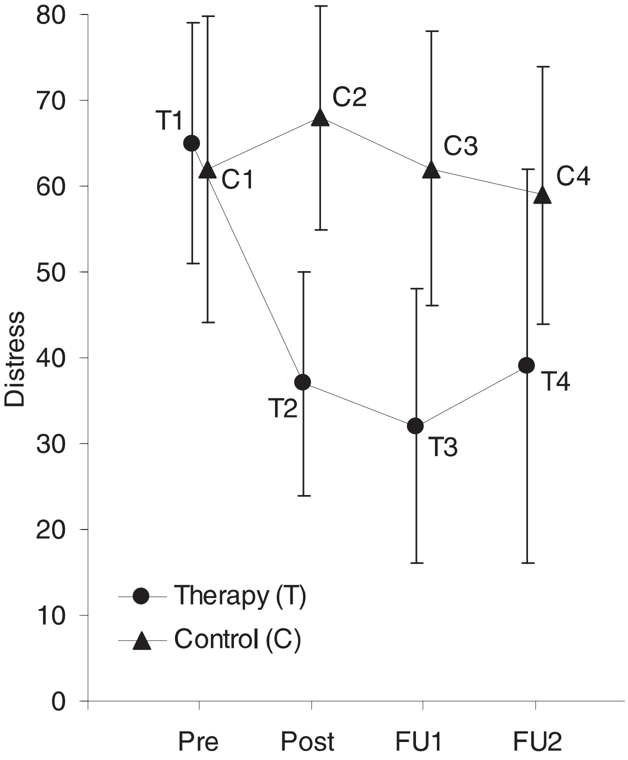

Figure 3. Means and 95% CIs for a fictitious experiment comparing a therapy group (T) that received treatment intended to reduce distress, and a waiting-list control group (C). Levels of distress were measured at pretreatment (Pre), posttreatment (Post), 1 month posttreatment (FU1), and 6 months posttreatment (FU2). The groups (T vs. C) is a between-subjects variable, and testing time is a repeated-measure variable.

or as an interval that specifies a set of plausible values for the population mean.

Figure 3 raises an issue vital for the correct interpretation of CIs: the importance of independent variable (IV) type. If means are independent, it is justifiable to compare the CIs on any two means and to consider the extent of any overlap of the two intervals: The CI widths are relevant for assessment of the difference between means. If there is little or no overlap, the *p* value for a comparison of the means is small, and we might conclude there is reasonable evidence of a difference between the two population means. By contrast, if the IV is a repeated measure, or there is matching, the CIs on the means may *not* be used to assess the difference between the means: The CI widths are quite *irrelevant* for assessment of the difference between means. To assess this difference, the CI on the paired *differences* is needed.

This distinction corresponds to that between the independent groups *t* test and the *t* test for paired data: The denominator for *t* for independent groups is based on the *SD* within each group, just as the CIs on the two means are based on those two *SD*s. Therefore, the two CIs present information relevant to assessment of the difference between means. By contrast, the denominator for repeated measure *t* is based on the *SD* of the paired differences, and so the mean difference must be assessed by the CI on these paired differences, and not by the CIs on the separate means. The higher the correlation between the two measures, the shorter is the CI on the differences: For given CIs on the separate means, the width of the CI on the paired differences may be anything from about twice as wide down to virtually zero, depending on that correlation. It is, thus, essential to be certain about IV type before considering the pattern of means and CI widths, and what conclusions are justified.

Cumming and Finch (2005) discussed and illustrated that distinction, and suggested a rule of eye, or approximate guideline, that can be used to assess overlap of 95% CIs on independent means: If the overlap of two such CIs is less than about half the average MOE of the two intervals, the *p* value for a comparison of the two means is less than .05. If the overlap is zero – the two intervals just touch end to end – or there is a gap, then *p* is less than .01. In Figure 3 the group IV (Therapy vs. Control) is independent and so the rule of eye can be used to assess the Therapy vs. Control comparison at any of the four testing times. Considering each of these comparisons separately, at Posttest the gap between the two CIs signals a *p* value distinctly less than .01; the small overlap (less than half the average MOE) at Follow-up 1 signals $p < .05$; and the considerable overlap at Follow-up 2 signals *p* distinctly greater than .05.

By contrast, the testing time IV is a repeated measure, and so the CIs displayed in Figure 3 may not be used to assess any within-group difference, for example Follow-up 1 vs. Follow-up 2 for the Therapy group, for which the appropriate repeated-measure test may give $p < .05$.

We should note that by discussing *p* values here we are not backing away from our earlier severe criticism of them, and are not advocating *p* values as the best way to interpret CIs on a pattern of means. Even so, as Cumming and Finch (2005) stated, it may be useful to discuss the relation between *p* values and CIs as one way to think about CIs that may bridge between familiar *p* values and less familiar CIs. The essential point about type of IV does not require any mention of *p* values: CIs on the independent means may be used to assess the difference, but the separate CIs on repeated-measure means convey information that is not relevant for assessing the mean difference.

Figure 3 presents fictitious data but the impression that, in general, the CIs are very wide, and that the experiment gave imprecise estimates of means and differences, is all too typical of psychology. Cohen (1994) famously stated: "I suspect that the main reason they [CIs] are not reported is that they are so embarrassingly large!" (p. 1002). We should not, however shoot the messenger and criticize CIs for their embarrassing width: CIs report accurately the extent of uncertainty in data. The problem is, rather, that a *p* value that leads us to reject the null, and conclude there is an effect, too easily seduces us into thinking we have nearcertainty and to overlook the true extent of remaining uncertainty.

Our topic is CIs, and all error bars we display are 95% CIs, but it is essential to keep in mind that the same error-bar graphic is, unfortunately, used to show various other quantities, including *SE* and, less often, *SD*. *SE* bars can, in many but not all situations, be interpreted as giving, approximately, 68% CIs (Cumming & Finch, 2005). However, it is CIs, unlike other options, that give inferential information and so CIs should be preferred whenever the interest is in inference, which means almost always. Also, it is 95% CIs that are most common and so, for consistency, it is best to use 95% CIs unless there are strong reasons to use a different level of confidence in a particular situation.

CIs for Proportions

Proportions lie between 0 and 1, and so the CI on a sample proportion is, in general, asymmetric because the 0 and 1 limits constrain the limits of the CI. The Appendix outlines a recommended method to calculate 95% CIs for proportions. It is an approximate method, but it is a close approximation and has desirable properties, including giving good CIs even for sample proportions of 0 and 1. Figure 4 displays 95% CIs for a range of proportions, for sample sizes of 30, 100, and 300. Only for a proportion of .5 are CIs symmetric, and the degree of asymmetry increases for proportions closer to 0 or 1. The intervals may seem surprisingly wide, and only narrow for very large samples, thus, illustrating again our tendency to underestimate the extent of sampling variability.

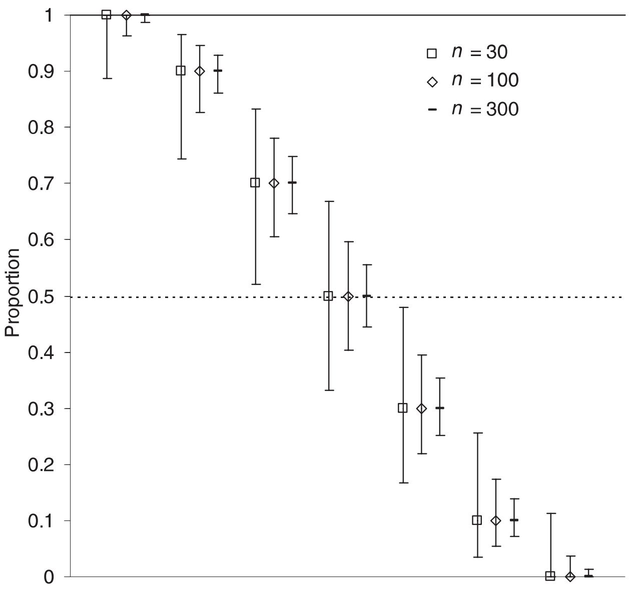

Figure 4. Sample proportions and 95% CIs for a range of proportion values, including 0 and 1, for sample sizes of 30, 100, and 300. Note the asymmetry of the intervals, especially for proportions near 0 and 1.

CIs for Correlations

Values of the Pearson correlation *r* are constrained to lie between -1 and 1 and, therefore, CIs for the population correlation *r* must also lie entirely within that interval. We should, therefore, expect that CIs on *r* are asymmetric, and indeed they are, for any value of *r* other than zero. CIs on *r* are based on Fisher's *r* to *z* transformation, as the Appendix explains. Figure 5 shows the 95% CIs for selected values of *r*, for sample sizes of 30, 100, and 300. It shows that the arm closer to zero is longer, and that the extent of asymmetry increases markedly as *r* approaches either limiting value of –1 or 1. It also may suggest that the CIs are, in general, surprisingly wide, and only become narrow for very large *n*: Again we tend to underestimate the extent of uncertainty.

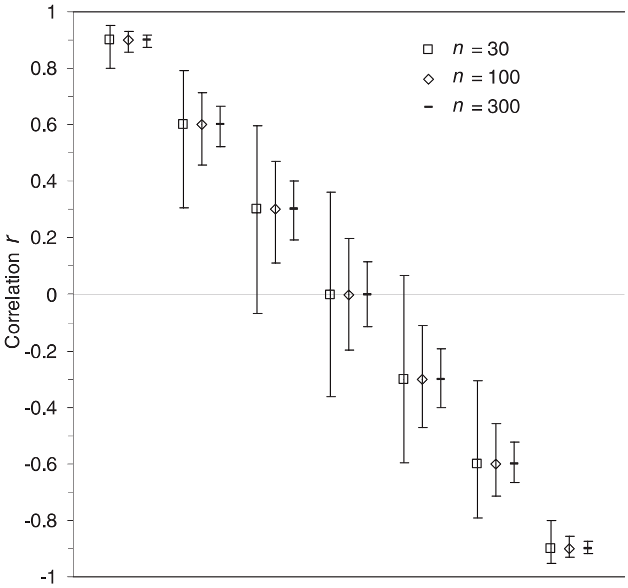

Figure 5. Sample Pearson correlations and their 95% CIs for a range of *r* values, for sample sizes of 30, 100, and 300. Note the asymmetry of the intervals, especially for correlations near –1 and 1.

CIs for Standardized ESs

Rosnow and Rosenthal (2009) introduce standardized ESs, and discuss the case of two independent groups. A standardized ES is the difference between the sample group means, divided by some *SD*. Choosing which *SD* to use can be difficult, so we discuss options here. If the population σ is known, and is the same for both groups, then the standardized ES is Cohen's *d* and is simply a *z* score. This situation may arise if, for example, the dependent variable is the score on an IQ test that is well-standardized to have σ = 15 in a reference population, and we judge it appropriate to use that *SD* for our research. In that case, we can find

CIs by using the familiar formulas for means as discussed above.

More commonly, the *SD* used to calculate the standardized ES is an estimate calculated from the data. If the *SD* of the control group is used, the ES is Glass's *D*; this may be the appropriate ES to use if a treatment changes the *SD*, as well as possibly the mean, and so it is unjustified to assume homogeneity of variance. If homogeneity of variance can be assumed and the pooled sample *SD* is used, the ES is Hedges's *g*. Cohen's *d* calculated from data, again assuming homogeneity of variance, is given by Rosnow and Rosenthal (2009) as $d = g$v$[(n_1 + n_2)/(n_1 + n_2 - 2)]$, where the two groups have sizes n_1 and n_2. Unfortunately, use of terms and symbols is not consistent in the literature, and the label Cohen's *d* is often used generically for any standardized mean (or mean difference). Cumming and Finch (2001), for example, presented analyses of *g*, but referred to it as *d*. It is, therefore, essential for authors to state the formula being used and, in particular, to make clear what *SD* is being used to accomplish the standardization.

If, as usual, the standardization *SD* is calculated from the data it is essential to appreciate that this *SD* contains sampling error. If the experiment were repeated, the *SD* would, no doubt, be different. The standardized ES is, thus, expressed in *SD* units that apply just to a particular data set: It is an ES with a "rubber ruler." A given value of *g* or *d*, 0.5 for example, reflects both the difference between sample means and the pooled within-group *SD* for that data set. Two experiments with the same *g* or *d* may have different mean differences – different ESs expressed in the original measurement units – if the two *SD* estimates are also different. Conversely, two different values of *g* or *d* may reflect the same mean difference – the same original-units ES – but merely sampling variation in the pooled *SD*.

This difficulty of standardized ESs has led some scholars, especially from medicine, to argue that standardized ESs are fundamentally flawed and should not be used. Greenland, Schlesselman, and Criqui (1986) argued that the standard unit is deceptive:

> The "standard unit" at issue here is merely the sample standard deviation of the study, and this quantity in no way conforms to ordinary English or scientific concepts of "standard unit": After all, the standard deviation can vary dramatically upon changing the study design, the variable under discussion, or the target population. . . . We think it is simply misleading to term something a "standard unit" when it is in fact a high variable quantity. (p. 207–208)

Medicine may be able to avoid such ESs because, in many cases, a single measure is used by all researchers measuring a given variable – blood pressure, for example, is universally measured in mm of mercury. In psychology, by contrast, it is common for a variety of instruments to be used to measure a single variable, anxiety for example. Standardizing those diverse anxiety scores appears to be necessary to permit direct comparison of different studies, or to meta-analyze a set of studies. Greenland (1998), to his credit, considered the comparability of results expressed in different measures, and suggested that perhaps benchmarks of biological or clinical importance could be established for each measure, and used as the basis for comparison.

The contribution of sampling variability to both the numerator (the difference between sample means) and the denominator (the pooled within-group *SD*) also complicates the calculation of CIs. Hedges's *g* is probably the most widely-used standardized ES for means, even if it is sometimes referred to, incorrectly, as Cohen's *d*, and so we consider here CIs only for *g*. Cumming and Finch (2001) and Smithson (2003) explained how accurate calculation of CIs for *g* requires use of the noncentral *t* distribution, and an iterative computer algorithm. Smithson provides scripts to calculate such CIs using SPSS, SAS, or R (see http://psychology.anu.edu.au/people/smithson/details/CIstuff/CI.html), and the ESCI software calculates CIs for *g* for simple cases (see www.latrobe.edu.au/psy/esci). Cumming and Finch, and Smithson, also explained that CIs for *g*, based on noncentral *t*, are not in general symmetric.

Rosenthal (1994, p. 238) gave this expression as an approximation for the *SD* of *g*:

$$\sqrt{\frac{N}{n_1 n_2} + g^2\left(\frac{1}{2(df)}\right)}$$

The Appendix explains how to use this to calculate approximate CIs for *g*. Figures 6 and 7 compare 95% CIs for *g* calculated using this approximation against accurate intervals based on noncentral *t*. Figure 6 compares the upper and lower limits of the approximate and accurate intervals for the case where $n_1 = n_2 = 5$. Figure 7 shows the error in the length of the upper and lower arms of the approximate intervals, expressed as a percentage of the length of the arms of accurate intervals. It shows this error when the sample sizes – always the same for the two samples – take various values from 5 to 100. The approximate intervals are always symmetric, whereas the accurate intervals are in general asymmetric, with the upper arm a little shorter than the lower arm when $g > 0$. The differences between the upper and lower arm errors in Figure 7, thus, reflect the small asymmetry of the accurate CIs.

The figures show values for *g* between 0 and 4, but note that most values of *g* arising in practice usually lie within about ± 1.5. Intervals and curves are symmetric about zero, and so the equivalent curves for $g < 0$ need not be displayed. Figure 7 shows that the approximation is generally excellent: Even for very small samples ($n_1 = n_2 = 5$) the arm length error is rarely greater than 3%, and usually less. For samples of at least 20 the error is usually no more than 1%. Exploration of cases with unequal sample sizes gives comparable conclusions.

Grissom and Kim (2005, chapter 3) and Hedges and Ol-

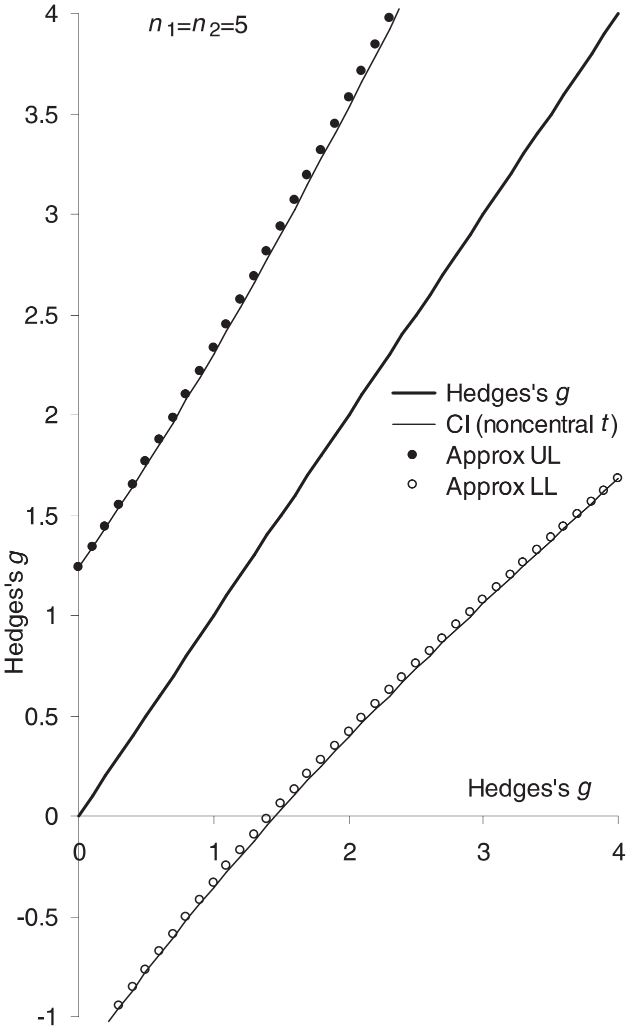

Figure 6. A comparison of 95% CIs for Hedges's *g* (heavy line), when $n_1 = n_2 = 5$. Accurate intervals are based on noncentral *t*, and approximate intervals are based on Rosenthal's approximation as described in the text. The fine lines mark the upper (UL) and lower (LL) limits of the accurate CIs, and the dots mark the upper (filled dots) and lower (open dots) of the approximate CIs, for values of *g* between 0 and 4. For larger samples, the approximate and accurate intervals are almost indistinguishable at the scale of the figure.

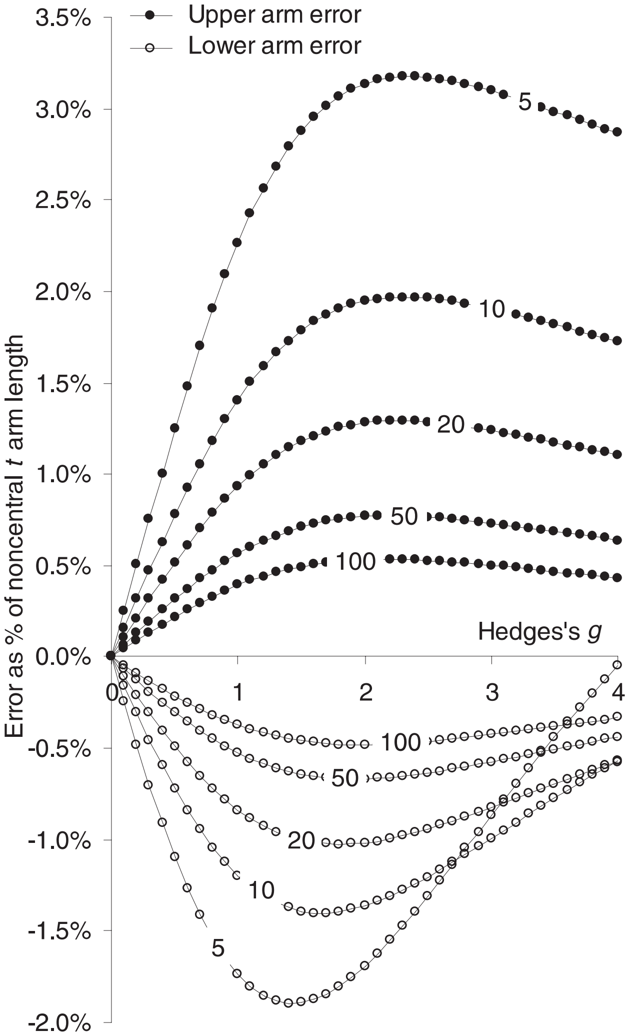

Figure 7. The error in upper and lower arm length of approximate 95% CIs for Hedges's *g* calculated using Rosenthal's approximation (see text), expressed as the percentage of the arm length of 95% CIs based on noncentral *t*, for values of *g* between 0 and 4. The numbers on the curves give the size of each of the two samples being compared, which are in every case equal. The solid dots give the percentage error for the upper arms. In general, the approximation gives upper arms that are a little long. The open circles are for lower arms and show that these are generally a little short. The two curves marked "5" correspond to the comparison shown in Figure 6.

kin (1985, chapter 5) discussed further the issue of CIs for standardized ESs. They gave approximate formulas for removing the small bias that *g* has as an estimate of the population standardized ES at low sample sizes. Grissom and Kim also considered cases in which the assumptions of normal populations and homogeneity of variance may not be warranted.

Our recommendation is that standardized ESs need to be used with care, and with careful appreciation that the standardization *SD* is subject to sampling error. Nevertheless, they are important for psychology, and should be reported and discussed where appropriate. In many cases, Hedges's *g* is likely to be the ES of choice. Interpretation of *g* or any other standardized ESs, as of any other ES, is a matter for informed judgment by the researcher in a particular research context. As mentioned above, software is now available for calculating accurate CIs, based on noncentral *t*, but the approximation discussed above gives an excellent and easily-used method of calculating CIs for *g*.

Conclusions

Throughout the half century of NHST dominance in psychology, distinguished scholars have published cogent critiques of the logical and philosophical foundations of NHST, of the way it is used in psychology, and of the limitations it imposes on the effectiveness of research. Kline (2004, chapter 3) provided an excellent overview. Few defences of NHST have been published, and almost all of these have admitted deficiencies of inference based on *p* values. Even so, NHST inertia has been overwhelming, and the technique continues to dominate.

We argue that estimation based on CIs should be used wherever possible as the primary way for data to be analyzed, and as the basis for interpretation of experimental results. Psychologists can immediately use CIs alongside NHST, then, as familiarity with estimation grows, it can progressively come to replace NHST.

CIs require care in interpretation, and an appreciation of the notions of replication and the infinite set of possible outcomes of an experiment, which provide the framework for understanding the level of confidence, usually 95%. It can also be useful to think of a CI in terms of the distribution of likelihood for various possible values for μ, the population parameter being estimated, as illustrated in Figure 2. Additionally, it can be useful to think of a CI as a prediction interval for future replication means. We hope these approaches can help psychologists appreciate the richness of information CIs provide, and the encouragement CIs give to ask questions requiring quantitative answers. Indeed, CIs offer better answers to better questions.

Acknowledgments

This research was supported by the Australian Research Council.

References

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

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

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

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen. *Psychological Science, 16*, 1002–1004.

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

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3.*

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A. et l. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230–232.

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*, 532–574.

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

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11*, 217–227.

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

Greenland, S. (1998). Meta-analysis. In K.J. Rothman & S. Greenland (Eds.), *Modern epidemiology* (2nd ed., pp. 643–673). Philadelphia, PA: Lippincott-Raven.

Greenland, S., Schlesselman, J., & Criqui, M. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology, 123*, 203–208.

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

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Killeen, P.R. (2005). An alternative to null hypothesis significance tests. *Psychological Science, 16*, 345–353.

Killeen, P.R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review, 13*, 549–562.

Killeen, P.R. (2008). Replication statistics. In J.W. Osborne (Ed.), *Best practice in quantitative methods* (pp. 103–124). Thousand Oaks, CA: Sage.

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

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

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–171.

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

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147–163.

Newcombe, R.G., & Altman, D.G. (2000). Proportions and their differences. In D.G. Altman, D. Machin, T.N. Bryant, & M.J. Gardner (Eds.), *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed., pp. 45–56). London: British Medical Journal Books.

Rosenthal, R. (1994). Parametric measures of effect size. In H.

Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosnow, R.L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Zeitschrift für Psychologie / Journal of Psychology, 217,* xxx–xxx.

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

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

Geoff Cumming/Fiona Fidler

School of Psychological Science
La Trobe University
Victoria
Australia 3086
Tel. +61 3 9-479-2820
Fax +61 3 9-479-1956
E-mail g.cumming@latrobe.edu.au / f.fidler@latrobe.edu.au

Appendix

Calculation of Confidence Intervals

CIs for Means

Assuming normally distributed populations, the 95% CI for μ, calculated for a sample with mean *M* and *SD*s is $[M - t_n s/\text{v}n, M + t_n s/\text{v}n]$, where t_n is the 95% critical value for *t*, for n = $n - 1$ degrees of freedom. This CI is symmetric about *M*.

CIs for Proportions

Following Newcombe and Altman (2000), consider the proportion $p = r/n$, where *n* is the total number of items of interest, and *r* of these have some particular feature. The proportion lacking this feature is $q = 1 - p$. Calculate $A = 2r + z^2$ and $B = z\text{v}(z^2 + 4rq)$ and $C = 2(n + z^2)$, where *z* is 1.96 for 95% CIs. Then the 95% CI for the population proportion is $[(A - B)/C\ (A + B)/C]$, an interval that is not in general symmetric about *p*.

For the difference between two independent proportions, p_1 and p_2, let $D = p_1 - p_2$. Use the method of the previous paragraph to find the 95% CIs for p_1 to be $[l_1, u_1]$ and p_2 to be $[l_2, u_2]$. Then the 95% CI for the difference between the population proportions is $[D - \text{v}((p_1 - l_1)^2 + (u_2 - p_2)^2), D + \text{v}((p_2 - l_2)^2 + (u_1 - p_1)^2)]$. This interval is not in general symmetric about *D*.

CIs for Correlations

To calculate CIs for the Pearson correlation, use Fisher's *r*-to-*z* transformation, calculate CIs for *z*, then back transform the limits of this CI to obtain the CI on *r*. The *r*-to-*z* transformation is $z = (1/2)\log_e((1 + r)/)1 - r))$, which as a Microsoft Excel function is $z = 0.5$*LN$((1 + r)/(1 - r))$. The SE of *z* is $\text{SE}_z = 1/\text{v}(n - 3)$, and so the 95% CI for *z* is $[z - 1.96\text{SE}_z, z + 1.96\text{SE}_z]$. To back transform the interval, substitute in turn its two limits as *z* in $r = (\exp(2z) - 1)/(\exp(2z) + 1)$, which as an Excel function is *r* = (EXP(2*z)–1)/(EXP(2*z) + 1). The CI is not in general symmetric about *r*.

CIs for Standardized Effect Sizes

To calculate CIs for Hedges's *g*, the first option is to use software, as described in the main text, to calculate accurate intervals based on the noncentral *t* distribution. The second option is to calculate the interval $[g - z_C SD, g + z_C SD]$, where z_C is the critical value of *z* for confidence level *C* ($z_C = 1.96$ for $C = 95$), and *SD* is Rosenthal's approximation to the *SD* of *g*, which is given in the main text. Rosnow and Rosenthal (2009) give an approximation to the *SD* of Cohen's *d*, and this can be used to calculate CIs for *d*, although as they explain a critical value of *t*, not *z*, is used.