Building confidence in confidence intervals

Graham D. Smith and **Peter E. Morris** encourage you to rely less on significance tests

'If you want to teach people a new way of thinking, don't bother trying to teach them. Instead, give them a tool, the use of which will lead to new ways of thinking.

Richard Buckminster Fuller

Statistics are tools to aid thinking. But a misused thinking tool can mislead.

We psychologists are frequently misled by statistics, albeit unwittingly (e.g. Coulson et al., 2010; Hoekstra et al., 2014). These misunderstandings contribute to a major problem where many effects reported in our journals cannot be replicated (Ioannidis, 2005; Kühberger et al., 2014).

This article aims to help readers reach sound conclusions about data, use better tools for the job and avoid some common misunderstandings. We encourage you to rely much more on confidence intervals and less on significance tests. If you do, then you will see research findings in a new light, enriching your understanding of psychological evidence.

We start by exploring some of the pitfalls of significance testing. Next, we introduce an alternative approach to inferential statistics based on confidence intervals of effect size. Then we give a series of examples demonstrating that this alternative approach is more revealing than significance testing. Finally, we reflect on why confidence intervals may have been overlooked.

Towards sharper tools

It is said that bad workers blame their tools, but it is worse to misuse them; like holding a cricket bat back-to-front, or knitting a cardigan using a pair of screwdrivers! Significance tests (i.e. statistical procedures that generate *p* values) are almost ubiquitous in quantitative psychology. Yet these tools are frequently misused. *P* values are sometimes taken to be: (i) a valid estimate of the magnitude of effects; (ii) the probability that the null hypothesis is true; (iii) the probability of replicating a result; or (iv) an indication of the theoretical or practical significance of results. Yet each of these interpretations is false (Nickerson, 2000). Typically, significance testing is used to determine the probability of obtaining the observed effect if a null hypothesis of zero difference or zero correlation were true. More useful non-nil null hypotheses can be employed but in practice rarely are. Many statistically savvy commentators, journal editors and psychological societies have concluded that we ought not to depend so heavily upon significance testing (APA, 2010; Cohen, 1994; Cumming, 2014; Nickerson, 2000).

Psychologists need to employ more ergonomic statistical tools; ones that are not so easily misused or misinterpreted. Confidence intervals (CIs) are the statistical equivalent of the ratchet screwdriver and the non-stick frying pan. Like significance testing, CIs are inferential statistics in that they enable us to draw conclusions regarding hypotheses about populations. However, they help compensate for our limited abilities so that we avoid many of the mistakes that significance tests encourage (Coulson et al., 2010).

How does one use CIs to evaluate a potential solution to an applied problem, or to test the prediction of a theory? Here is a two-step quick-start guide.

Step 1: Calculate a point estimate of the population effect size.

To evaluate a hypothesis the magnitude and direction of an effect must be quantified. The effect might simply be the difference between two means. Or the mean difference could be standardised by dividing it by the scores' pooled standard deviation (SD), yielding Cohen's *d*. Pearson's correlation coefficient is an effect-size measure of the relationship between two variables. There are many other effect-size measures designed for a host of situations (see Fritz et al., 2012; Morris & Fritz, 2013b).

An effect size is not just a useful description of a sample; it guides us to conclusions about the population from which the sample is drawn. Many effect-size measures are designed to give unbiased point estimates of the true size of the effect in the population. In other words, the sample's effect size tells us the most plausible value of the population's effect size.

Step 2: Calculate the likely range of the population effect size.

Whilst a point estimate of the population effect size is useful, we need to remember that the actual population effect size could be among a range of values either side of the sample effect size. It is less and less plausible that the population effect size is at values farther and farther away. The distribution of plausibility (i.e. relative likelihood) of various values of the population effect size, under the parametric assumptions, is shown in Figure 1.

It would be useful to know how far

references

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

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

Armitage, C.J. & Talibudeen, L. (2010). Test of a brief theory of planned behaviour-based intervention to promote adolescent safe sex intentions. *British Journal of Psychology, 101*(1), 155–172.

Baguley, T. (2012). Can we be confident in our statistics? *The Psychologist, 25*, 128–129.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* Hillsdale, NJ: Lawrence Erlbaum.

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

Coulson, M., Healey, M., Fidler, F. & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1*(26), 1–9.

Cumming, G. (2012). *Understanding the new statistics.* New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29.

Fidler, F. & Loftus, G.R., (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217*, 27–37.

Fritz, C.O., Morris, P.E. & Richler, J.J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology:*

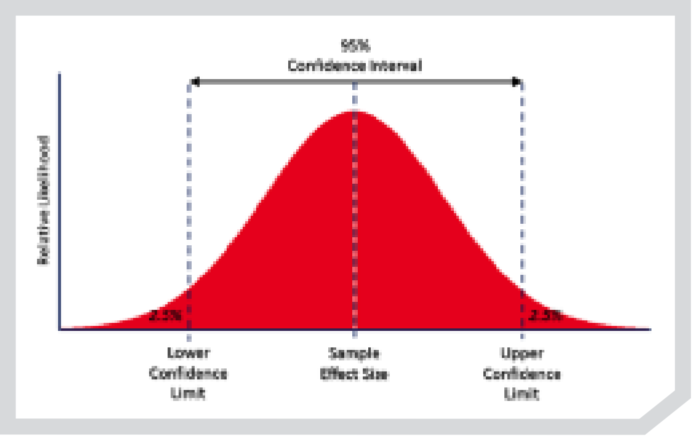

Figure 1. The plausibility of potential locations of the (unknown) population effect size. The most plausible value of the population effect size is the same as the sample effect size. It is very likely that the 95 per cent confidence interval has captured the population effect size, but not certain.

away from the population effect size the point estimate might plausibly be. What range encompasses 95 per cent of the most likely effect size values? This is just what a 95 per cent CI tells us. It helps us to visualise the distribution of plausibility among potential population effect sizes. A value near the sample effect size is about seven times more plausible than a value close to the 95 per cent confidence limits (Cumming, 2012). Only rarely (5 per cent of the time for 95 per cent CIs) will the CI turn out not to have captured the population effect size. CIs indicate the precision of estimates of effects.

Together, point and interval estimates of population effect sizes allow us to draw conclusions about almost any research question. There are CIs for nearly every situation that there are tests of significance (see Altman et al., 2000; Fidler & Loftus, 2009; Newcombe, 2013). The CI methods described later require the parametric assumptions but non-parametric CIs exist too.

CIs deal directly with research hypotheses in that they allow hypotheses about the many potential values of an effect to be evaluated together (Cumming, 2012). From a single CI, one can know whether the plausible range of effect sizes includes negligible, small, large, positive, negative and humongous values. A single significance test invites a conclusion about only one null hypothesis that typically is tangential to the research hypothesis. Multiple significant tests with different null hypotheses would be required to match the versatility of CIs. Furthermore, CIs ensure proper attention is paid to the magnitude of effects and thereby discourage many misinterpretations that are promoted by significance testing.

Now, let us build confidence in CIs through a series of examples of their interpretation, compared with conclusions from significance testing. We have chosen to use imaginary data for some of the examples. This is not because real examples are rare, but because we do not want to imply criticism of the few authors that we might select.

Statistically significant yet negligible

In his meta-analysis of personality variables and academic achievement, Poropat (2009) found statistically significant correlations (both *p* values were < .001) between academic achievement and both extraversion and stability. Described solely in terms of statistical significance, this sounds as if there are important relationships between academic achievement and these two personality dimensions. Maybe this should be being taken into account by universities in their selection and treatment of students? However, Poropat was not misled by the significant *p* values, because the correlation coefficients of extraversion and stability with academic achievement that he found were –.01 and .01 respectively. If we estimate, by squaring the correlations, the variance of academic achievement predicted by extraversion or stability we see in each case the figure is only 0.01 per cent. The correlations may be statistically significant but the relationships are virtually non-existent. They are significant not because the effects are substantial but because the meta-analysis has accumulated sample sizes of over 59,000. If a sample is large enough then even the most trivial effects are statistically significant (Nickerson, 2000). Poropat rightly overlooked the *p* values and drew conclusions from effect sizes. Furthermore, we calculate the 95 per cent CIs for the correlations of academic achievement with extraversion and stability to range between .00 and –.02, or .00 and .02 respectively. In other words, it is reasonable to conclude that the actual population values for these correlations are negligibly tiny.

Significance testing cannot give grounds for believing a null hypothesis, partly because absence of evidence is not evidence of absence. Furthermore, non-significant results do not imply that observed effects are negligible, as this can equally arise through low power. Yet, CIs can demonstrate that, under certain conditions, even a significant effect is so small as to be not worthy of our attention.

Statistically significant yet potentially not substantial

We have just seen how the range of likely values indicated by the CIs is very small if the sample size is very large. However, with the sort of sample size common in

General, 141, 2–18.
Grissom, R.J. & Kim, J.J. (2011). *Effect sizes for research: A broad practical approach* (2nd edn). New York: Psychology Press.
Hoekstra, R., Finch, S., Hiers, H.A.L. & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*(6), 1033–1037.
Hoekstra, R., Morey, R.D., Rouder, J.N. & Wagenmakers, E.J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164. doi:10.3758/s13423-013-0572-3.
Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi: 10.1371/journal.pmed.0020124.
Kühberger, A., Fritz, A. & Scherndl, T. (2014). Publication bias in psychology. *PLoS ONE, 9(9)*. doi:10.1371/journal.pone.0105825.
Morris, P.E. & Fritz, C.O. (2013a). Effect sizes in memory research. *Memory, 21*(7), 832–842.
Morris, P.E. & Fritz, C.O. (2013b). Why are effect sizes still neglected? *The Psychologist, 26*, 580–583.
Morris, P.E. & Fritz, C.O. (2014). *The challenge of the Psychonomic Society guidelines on statistical issues (2012).* Poster presented at the Psychonomic Society Annual Conference, Long Beach, California.
Morris, P.E., Fritz, C.O. & Buck, S. (2004). The name game: Acceptability, bonus information and group size. *Applied Cognitive Psychology, 18*, 89–104.
Newcombe, R.G. (2013). *Confidence intervals for proportions and related measures of effect size*. Boca Raton, FL: CRC Press.

many experiments, the CI of even a significant population effect may be quite wide. Typically, researchers fail to avail themselves of this useful and sobering extra information. For example, one of us co-authored a paper (Morris et al., 2004) reporting an experiment in which participants played two versions of a game intended to promote recall of people's names. The elaborate name game was rated as being significantly more fun to play than the simple name game; *t* (214) = 1.79, *p* = .04 (one-tailed). This finding was taken as support for a minor prediction.

However, secondary analysis using the CI of an effect size goes beyond the *p* value. Cohen's *d*, calculated from *t* and the sample size (Fritz et al., 2012), is 0.24. This can be referred to as a small effect (Cohen, 1988) although categorising effect sizes ought to be undertaken with caution (see Morris & Fritz, 2013a). A one-tailed 90 per cent CI of *d* is consistent with the finding of a significant difference. However, suppose we were instead interested in the range within which the population mean for this comparison might lie if there were no prediction. Cumming's (2012) ESCI software indicates that the 95 per cent CI of *d* is –0.04 to 0.52. In other words, it is reasonable to conclude that the population effect size is somewhere between a negligible effect and a medium-sized positive effect. Only further research with much larger samples and possibly tighter control of variance in the design can answer whether Cohen's *d* really is substantial.

Fortunately, this particular comparison was not central to the conclusions of the paper, but it does illustrate how CIs can lead researchers to more appropriately nuanced conclusions about significant results. Any effect size must always be interpreted in terms of its practical importance. For example, a small effect size can still be a valuable discovery when it relates to life and death health issues.

"Significance testing encourages misleading black-and-white thinking "

Not statistically significant yet potentially substantial

Imagine your research compares the language development of girls and boys. You find that a sample of 16 girls (*M* = 57.6, SD = 21.4) scored higher than a sample of 16 boys (*M* = 46.7, SD = 18.5) on a measure of verbal fluency. But an independent measures *t* test shows the difference is not significant, *t* (30) = 1.54, *p* = .13.

What should you conclude? You might be tempted to infer that the null hypothesis is probably true. Concluding that a null hypothesis is likely to be true because an effect is not statistically significant is a mistake frequently seen in the literature (Hoekstra et al., 2006). So perhaps you would drop this line of inquiry and try something else with a better chance of being successful. Would you consider a *p* value of .13 close enough to the .05 criterion to justify the time and cost of collecting more data? On its own, the *p* value does not give you the information to decide. But a CI does.

The mean difference of girls' and boys' verbal fluency scores (*M* = 10.9) translates to a Cohen's *d* of 0.56. The 95 per cent CI of Cohen's *d*, calculated from *d* and the sample sizes (Cumming, 2012; Grissom & Kim, 2011), is –0.14 to 1.27. Towards the lower limit, the population *d* could plausibly be between a very small negative value and negligible positive value, consistent with the finding of non-significance. But it would be wrong to discount the effect because of the possibility that it is negligibly small. Remember both extremes are equally likely. Towards the upper limit, the true population effect could just as easily be large or very large (Cohen's *d* of 1.27 accounts for 30 per cent of variance; Fritz, et al., 2012). We may not have precise enough an estimate of the population effect size to justify publication, yet we ought not to assume that the effect is negligible when it is much more plausibly medium-sized, large or very large. It is well worth attempting replication with a larger sample. The *p* value was in danger of encouraging us to abandon a potentially exciting line of inquiry.

Contradictory yet consistent?

Imagine two articles that report essentially the same independent measures experiment replicated by different researchers. The difference of means in Study A is 13.9 (SD = 20.0), *t* (38) = 2.20, *p* = .03. The difference of means in Study B is 9.1 (SD = 21.1), *t* (28) = 1.18, *p* = .25.

Are the findings consistent or contradictory? In one study the difference is statistically significant but in the other study it is not. Would you conclude that the findings are inconsistent? Might you then carefully examine the articles for methodological differences to explain the outcomes?

Perhaps you are being tempted into a mistake. Significance testing encourages misleading black-and-white thinking (Hoekstra et al., 2006; Nickerson, 2000). Statistical significance and non-significance do not equate to the existence and non-existence of an effect. A significant effect is not necessarily significantly greater than a non-significant effect (Baguley, 2012). Ought you to conclude that the findings complement each other? On their own, the *p* values cannot help you decide.

A clearer picture emerges when we look at CIs. The CI of the (unstandardised) difference of means is calculated easily from means, SDs and sample sizes (e.g. Altman et al., 2000). The 95 per cent CI for Study A is 1.11 to 26.71 and for Study B is –6.71 to 24.91. These intervals are largely overlapping, so it is quite plausible that there is little difference in the true size of effect in the two studies.

But do they overlap enough? It seems so. A meta-analysis combining the two studies using Cumming's (2012) ESCI software reveals the 95 per cent CI is 2.38 to 21.56 and the *p* value associated with the null hypothesis of zero difference is .01. You may be surprised that this *p* value is less than either of the *p* values of the original studies. However, this surprise would be misplaced; the evidence for statistical significance is accumulated because the study findings are largely consistent and the sample size is cumulative.

Using and misusing CIs

Sadly, few empirical articles report CIs and even fewer interpret the CIs they report. Morris and Fritz (2014) found that only 11 per cent of the 463 empirical papers published by the Psychonomic Society's journals in 2013 reported CIs, and very few interpreted these CIs.

Looking back through the last five years of the *British Journal of Psychology* we can find only one article that puts CIs at the heart of the analysis and

Nickerson, R.S. (2000). Null hypothesis significance testing. *Psychological Methods, 5*, 241–301.

Poropat, A.E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338.

interpretation: Armitage and Talibudeen (2010). They report the mean difference in acceptance of a safe-sex message between an attitude-change intervention and a control was 0.82, 95 per cent CI [0.57, 1.07]. Had they merely stated the means and a *p* value then we would not know how small or large the difference could plausibly be. For all we would have known, the population difference in message acceptance could just as easily have been somewhere between 0.07 and 1.57, a much less impressive finding. The authors also tell us that Cohen's *d* for the effect is 0.77. Were they preparing the paper today they could have used Cumming's (2012) ESCI to determine the CI of *d*. Our secondary analysis of their results suggest that the effect in the population is likely to be between medium and large; 95 per cent CI [0.48, 1.06]. This suggests that the change in attitude is not just potentially greater than zero, it is substantial. One might be able to make a case for the effect being large enough to make a practical difference in safe-sex behaviour.

If CIs are superior to significance testing, why are they reported so infrequently? It would be wrong to conclude from this neglect that CIs are not superior after all. Here are several potential explanations. First, psychologists may not know enough about CIs to realise their usefulness; the basics are not widely known. Many mistakenly believe that CIs are merely descriptive statistics, and even experienced researchers are unaware that CIs exist for Pearson's *r* and Cohen's *d*. Few statistical textbooks and statistical packages do sufficient justice to the CI approach. Second, one ought not to underestimate the normative influence of the literature. Results reporting may be a matter of habit. The lack of good examples such as Armitage and Talibudeen (2010) may mean that researchers fail to consider using CIs. A third reason why CIs may not be reported is that they often reveal a very wide range within which the population values may lie (Cohen, 1994). It is uncomfortable to be reminded of the imprecision of one's data, but trying to ignore that imprecision will not make it go away.

Although CIs are better tools than significance tests, they are not foolproof. Hoekstra and colleagues (2014) recently reported that experienced researchers hold misunderstandings about CIs. However, the misunderstandings may be due to lack of familiarity. Coulson and others (2010) found that because psychologists are relatively unfamiliar with CIs they frequently reinterpret them as two-tailed significance tests, by observing whether or not zero is captured within the interval. There is little point in merely using CIs in this way – it leads to the same conclusions as conventional significance testing. CIs, when interpreted appropriately, yield richer conclusions than significance tests; more nuanced and possibly in contradiction.

To see the real utility of CIs we need to go beyond the significance testing mindset. Only then will we beam new light on research findings. So take these powerful thinking tools out the box and give them a go!

Graham D. Smith *is a Senior Lecturer at the University of Northampton*
Graham.Smith@northampton.ac.uk

Peter E. Morris *is Emeritus Professor at the University of Lancaster and Honorary Visiting Professor at the University of Northampton*