

Interpreting Significance: The Differences Between Statistical Significance, Effect Size, and Practical Importance

Pawel Kalinowski, BBSc and Fiona Fidler, PhD

It is a common misconception that statistical significance indicates a large and/or important effect. In fact the three concepts—statistical significance, effect size, and practical importance—are distinct from one another and a favorable result on one dimension does not guarantee the same on any other. In this article, we explain these concepts and distinguish between them. Finally, we propose reporting confidence intervals as a step toward disambiguating these concepts.

Keywords: Confidence intervals; Effect size; Clinical significance; Misconceptions

Researchers pride themselves in reporting a “significant result” and null hypothesis significance testing (NHST) is overwhelmingly the most common statistical analysis in most social, life, and behavioral sciences.¹ For example, in 10 leading psychology journals, NHSTs were reported in more than 95% of articles.² An analysis of the last 10 issues of *Newborn and Infant Nursing Reviews* (NAINR) shows that this discipline is no exception. More than half (56%, 43/77) of articles published in 2007 to 2009 referred to the “significance” (or lack thereof) of results. The remainder were not empirical articles or reported descriptive statistics such as means only and did not provide any inferential tests. Unfortunately, NHST is the subject of great controversy, and many pitfalls and misconceptions have been associated with its use. In this article, we review four of the most common errors associated with this practice and recommend reporting confidence intervals (CIs) as a solution.

What is Statistical Significance?

Before we introduce our four common errors, let us revisit the definition of statistical significance. A result is deemed “significant” if the *P* value produced from a NHST is below a certain value, traditionally .05 or .01. *P* is a probability value, specifically the probability of incorrectly rejecting the null hypothesis when in fact it is true. In other words, it is the

probability of a “false positive.” Think of a pregnancy test: *P* is analogous to the probability of the test telling you are pregnant when in reality you are not. This is a useful probability value to know, but equally useful is the probability of a false negative. We would not trust a pregnancy test that did not report its false-negative rate, and yet, many researchers make the equivalent leap of faith every time we interpret a *P* value without statistical power! Statistical power measures the false-negative rate (power = 1 – false-negative rate). It is the chance of detecting a statistically significant *P* value, if the effect really exists. This effectively makes it a measure of how good a study is: was there ever a reasonable chance of finding a statistically significant result? Although we traditionally keep the false-positive error rate at .05 or .01, as mentioned above, false-negative error rates have historically gone unmeasured.^{3–5} More recent surveys of the statistical power of published research show little improvement.⁶ In fact, there were no reported tests of statistical power in our analysis of the last 10 NAINR journals.

The crucial message here is that a *P* value is a very limited piece of information, relating to false-positive error rates only. Statistical significance is merely a statement about the *P* value relative to an arbitrary cutoff, so it too relates only to false-positive errors. There is much more to know about a set of empirical data!

One particularly common problem is that statistical significance is used ambiguously. In our brief survey of NAINR, we found almost two thirds of articles (63%, 27/43) used the term *significant* in an ambiguous way. By this, we mean that it was impossible to tell from the reports whether the author meant to refer to the statistical significance or the practical importance of their result. The statement “interventions have a significant impact on the degree of growth” is an example of an ambiguous significance report. Is this author reporting on a specific statistical outcome or are they highlighting the practical importance of their result? Perhaps, they are assuming that the two are equivalent? Shortly, we will explain why statistical significance, effect size, and practical

From the School of Psychological Science, La Trobe University.

From the Environmental Science, School of Botany, University of Melbourne, Bundoora, Victoria, Australia.

Address correspondence to Pawel Kalinowski, BBSc, Psychological Science, La Trobe University, Bundoora, Victoria 3086, Australia.

E-mails: p.kalinowski@latrobe.edu.au, f.fidler@latrobe.edu.au.

URL: <http://www.latrobe.edu.au/psy/staff/fidler.html>.

© 2010 Elsevier Inc. All rights reserved.

1527-3369/09/1001-0341\$36.00/0

doi:10.1053/j.nainr.2009.12.007

importance are not equivalent. We argue that the best way to avoid this confusion is to replace NHST with effect size estimates and a measure of precision associated with those estimates, called CIs.

Four Common Mistakes Made When Interpreting Statistical Significance

None of the mistakes we outline here are specific to nursing; they are made just as often in psychology, ecology, economics, and virtually every other life and social sciences.⁷ We are definitely not singling out nursing, as our frequent references to psychology examples, should indicate.

P Values are Not Measures of Effect Size

A common mistake is to assume that if *P* is small (eg, <.05 or .01) then the effect must be large. This is not necessarily the case: *P* is a product of not only effect size but also sample size, so if the effect is small but the sample size very large, *P* will be small and therefore statistically significant. Conversely, if the effect size is large and the sample size small, *P* will also be small.

For example, consider the Steering Committee of the Physicians' Health Study Group study⁸ on the effect of low doses of aspirin in reducing heart attacks. In this study, 11 034 men were given an aspirin to take every 2 days, whereas 11 037 men were given a placebo. With such large samples, it is hardly surprising that the difference was statistically significant at *P* < .000001. Although the *P* value was tiny (and so very statistically significant) the effect size was also tiny ($r^2 = .0011$).

P can be made small by increasing the sample size or enforcing tighter constraints on conditions to reduce SD, both of which will increase statistical power. When statistical power is very high, statistical significance ensues. As power approaches 1, statistical significance becomes a guarantee. The belief that a small *P* equals large effect is sometimes referred to as the *magnitude fallacy*.⁹ The take home message is this: *P* is not a straight indicator of effect size because it conflates effect size with sample size.

There is only one way to determine the effect size of a study and that is to look at the effect size! This may be a standardized effect size measure (eg, Cohen's *d*), a correlation, a mean difference, or any other measure that meaningfully summarizes what went on in the study. The American Psychological Association strongly encourages reporting effect sizes in original units: "Effect sizes may be reported in original units... and are most easily understood in original units."¹⁰ In medicine, original units are also preferred.¹¹

Effect sizes—in whatever form they are gathered—are best reported with a measure of precision. We recommend CIs for this purpose. Confidence intervals provide an account of the uncertainty associated with the effect size measure. If the CI is wide, then the effect estimate is uncertain, or imprecise. If the CI is narrow, then it is more precise.

Effect Sizes are Not Measures of Importance

Above we explained that statistical significance is not a measure of effect size. Now we turn to a second common misconception, that is, that effect size is a measure of importance. It is a common mistake to think that because the effect is large, it must be important. Perhaps even more common is the reverse error—assuming a small effect is of no importance. Let us dispel this misconception clearly. A large effect will not necessarily be clinically important, and sometimes small effects are clinically important.

Let us return to our aspirin example. Recall that *P* < .000001, even though the correlation between aspirin consumption and instance of heart attack was tiny ($r^2 = .0011$). This small correlation is of great clinical importance. In fact, this effect was so important that study was immediately terminated. The data showed for every 100 people in the aspirin condition, 48.3 had a heart attack, whereas for every 100 people in the placebo condition, 51.7 had a heart attack. In other words, there were 3.4% less heart attacks in the aspirin condition; 3.4% might not sound like a lot but remember the total number of patients in the experiment was 22 000—3.4% of 22 000 is a big number! (It is about 7 less heart attacks for every 200 people!) It was therefore deemed unethical to deny the control group an easy treatment.

Statistical Significance Does Not Measure Importance

So far we have established that (1) *P* does not measure effect size and (2) that effect size does not measure importance. The third step is that statistical significance is not an index of importance. Recall from our introduction that *P* is simply the probability of rejecting the null hypothesis, when in fact it is true (the probability of a "false positive"). It does not provide any more information than this. We do not expect a pregnancy test to tell us how many weeks pregnant we are or how important the pregnancy is to us. Its job is simple yes or no diagnosis, within a certain false-positive and false-negative error rate. Similarly, NHST does not tell us effect size or clinical importance.

So what statistic does tell us about clinical importance? Most often, it is our own judgment. Think of the aspirin example again. How did we know that 7 of every 200 lives was an important number? It was not the *P* value that told us this nor was it the correlation. It was common sense! Importance, be it clinical, theoretical, or biological, is a decision to be made by the experimenter not by a statistical test.

There are some guidelines to help us make this decision. For example, a special issue of the *Journal of Consulting and Clinical Psychology* was devoted to the question of how to measure clinical importance. Kazdin¹² provides an overview of the concepts and measures discussed in the special issue and is a good starting point for information on clinical significance. Several of indexes detailed in individual articles in this issue will be relevant to nursing research. For example, Jacobson et al¹³ define clinical significance as "returning to normal functioning." A normative comparison can be reported as the proportion of

patients whose baseline measures (on a blood pressure, depression, or any other kind of index) return to within a normal range after some intervention. The cutoff points or criteria for what counts as “normal functioning” will vary with individual measures and sometimes will be a matter of researchers' judgment.

P is Not a Measure of Replicability

This is sometimes referred to as the “replication fallacy.” It is a belief that a statistically significant result at, for example, $P < .01$ means that there is a 99% chance that if the experiment was repeated, the results would be replicated. The classic study of Oakes¹⁴ showed 60% of psychology researchers held this misconception. In fact, a P value does not give any direct information about the chance of replication. More recent replications of this study show that misconceptions persist.^{15,16} What is the predictive power of a statistically significant result? For example, what does $P = .02$ tell you about the P value you would get in a direct replication? For a $P = .02$, there is an 80% chance the next P will be between .0003 and .295.¹⁷ That is a very wide interval—from very small to well over any reasonable statistical significance cut off! In other words, P tells us very little about replication. Simulations showed (2-tailed) P values ranged from 0.001 to .76 for a typical experimental design of 2 normal distributed groups with $N = 35$, an SD of 20 with an actual mean difference between the groups of Cohen's $\delta = 0.5$, a medium effect.¹⁷

What does this mean for nursing research? Given that replication is regarded as the gold standard of science, it is hardly surprising that people assume that it would be reported in research results. It is important to note that P values do not provide this information. Confidence in the use of certain nursing procedures that have previously reported small P values may be unwarranted. Rather than relying on a P value to inform the reader of the replicability of the data, it is better to use error bars, such as SD, standard error, or (preferably) CIs. With a 95% CI, we can be 83% sure that if the study was replicated, the results would fall within the upper and lower bound of the CI.¹⁸

There are many other misconceptions about P values, and even those who specialize in methodology are not immune to them. Haller and Krauss¹⁵ found that 100% of students, 90% of academic psychologists, and 80% of methodology instructors held at least one misconception of P values. For a more complete list, see Fidler⁷ and Kline.⁹

What is a CI?

A CI is an estimate of a population parameter (eg, mean difference) calculated from a sample data. The confidence level of a CI represents the percentage of intervals (eg, 95%) that would capture the mean or other population parameter if the experiment was replicated an infinite number of times. Values within a CI are a good bet for the population mean and those outside it are not. However, it is important to remember that the

Table 1. Number and Percentages of NAINR Articles (Overall) that Reported CIs, Measures of Variability or Mentioned Practical Importance

	CIs	SE, SD, or Other Error Bars	Practical Importance
No. of articles	3	6	7
%	4	8	9

distribution of a CI is not uniform. Cumming and Finch¹⁸ give a list of similar interpretations.

- (i) Values inside a CI are plausible values for μ . Values outside the CI are relatively implausible.
- (ii) We can be 95% confident that our CI includes μ .
- (iii) Our data are compatible with any value of μ within the CI but relatively incompatible with any value outside it.
- (iv) The lower limit is a likely lower bound estimate of the parameter; the upper limit a likely upper bound.¹⁸

Cumming and Finch¹⁸ also offered the following useful ways to think about CIs:

- (1) The mean of your sample is your best estimation for the population mean; values closer to the mean are relatively more likely to be the population mean compared to points further away from the sample mean.
- (2) The width of a CI at a given level of confidence indicates the precision of the estimate. A small interval is more precise than a wide interval. There are three things that affect the precision of our CI. The sample size (the larger the sample size, the more precise the CI), the SD of the data (larger SD of the data reduces precision of the data), and finally the confidence level (a 95% CI is wider than a 90% CI as it captures more data).

A CI is just one from an infinite sequence. When routinely reporting CIs with a 95% confidence level, you can expect, over a lifetime, that about 95% of those intervals will capture the population parameters. The confidence level represents the frequency that the estimates would capture the population parameter given the experiment was repeated many times. Fifty percent CIs would capture the population parameter half of the time and 95% CIs would capture the population parameter around 19 of 20 replications. Unfortunately, CIs were quite rare in our analysis of NAINR journals (Table 1) with only 4% (3/77) articles overall reporting a CI.

Benefits of CIs Over Statistical Significance

CIs give you all of the information that P values give and more! CIs combine information on effect size and precision and can be used to conduct significance tests if required (by determining whether the null value is in or out of the interval). Significance tests require the separate reporting of a P value,

effect size, and a statistical power calculation to provide equivalent information. Seeing all three pieces of information (P value, effect size, and power calculation) in a journal article is extremely rare. Even if the information was reported, integrating three separate pieces of information is more cognitively taxing than being presented with a CI.

The superiority of CIs over NHST for the communication of data was demonstrated in a study by Fidler,⁷ in which 79 final year bachelor or masters year students were presented with data from a fictional study. The data were from a low-powered study that did not achieve statistical significance despite the presence of an important medium-sized effect. When given with results in NHST format (P values, statistical power, effect size), only 61% of students correctly identified the presence of an important effect. However, when these students were presented with the same information in a CI format (95% CI around effect size), 87% of the students correctly identified the effect. The misconception that statistical nonsignificance means no effect (an extension of mistake 2) is not entirely absent when results are presented in CI format, but it is less frequent.

CIs are also a natural complement to meta-analysis and meta-analytic thinking. CIs can be used directly in a meta-analysis and facilitate meta-analytic thinking.¹⁸ The importance of replication in research is being increasingly emphasized, and the growth of meta-analytic databases such as that of the Cochrane Collaboration is testament to this. A CI provides the necessary information needed when conducting a meta-analysis. At the very least, a researcher can immediately

compare current data, for example, the effect size and precision of one treatment to the effect size and precision of a previous more common treatment.

Common Mistakes With CIs

Although CIs are much more informative and intuitive than NHST, they are not free from misconceptions of their own. Empirical studies show that CIs are sometimes misinterpreted by students and researchers.^{7,19,20} The following is a list of some common misconceptions to watch out for when using CIs in your own research.

Definitional misconceptions: Fidler⁷ found that some students believed that CIs are a descriptive, rather than inferential, statistic. For example, they mistakenly believed the CI to be an estimate of the sample mean rather than of the population mean. Other common mistakes in this category included interpreting the CI as a range of individual data scores something akin to a SD. Remember, a CI is an estimate of plausible values for the population mean. It is inferential, making it a very different statistic to the SD.

The overlap misconception: The overlap misconception is a belief that when comparing means for two independent groups, the means are statistically significantly different at a $P < .05$ when the 95% CIs around the means are just touching (Fig 1). In actuality, CIs for independent groups are at $p \approx .05$ when they overlap by about $\frac{1}{4}$.¹⁷

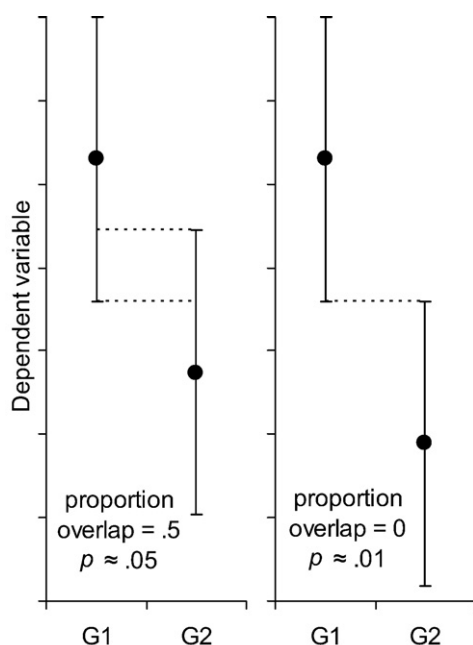


Fig 1. The CIs on the left overlap by about $\frac{1}{4}$, half the average margin of error, which corresponds to a P value of $\approx .05$. The CIs on the right are just touching. This corresponds to a P value of $\approx .01$ (Cumming and Finch¹⁸).

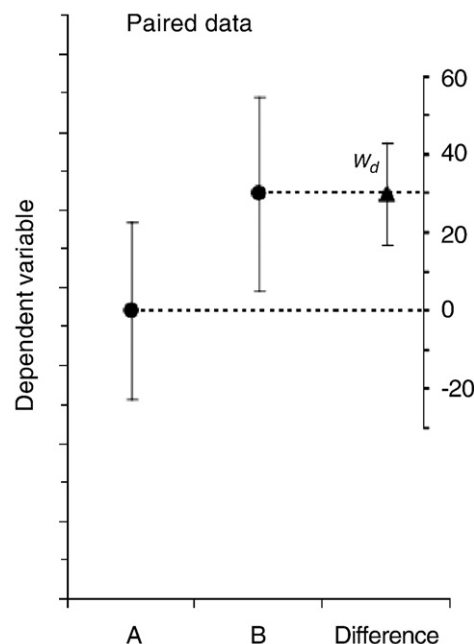


Fig 2. For paired or repeated measure designs a CI is placed on the mean difference. Confidence intervals on the group means will not give you a valid interpretation on the difference between the means.

Treating repeated measures such as independent means: When using a repeated measures or a paired design, do not compare CIs of the group means (Fig 1) as they use the SD of each persons' score relative to the mean. You will not get a meaningful interpretation. Instead, you want a CI around the mean difference (Fig 2) using the SD of the difference between each person's score for pretest and posttest.

The confidence level misconception: This is a belief that a 95% CI for an experiment has a 95% chance of capturing the population parameter for the next repeated experiment. This would only be true if the sample mean landed exactly on the population mean, which is unlikely. In fact, the average probability of the first 95% CI capturing the next sample mean is around 83%.¹⁹

Conclusion

The clinical importance of a result depends on much more than its statistical significance. The latter should not be used as indicator of the former. In fact, effect size, statistical significance, and practical importance are separate data characteristics and should each be measured and judged in their own right. We believe CIs help separate these important issues, making it much less likely to mistake one for the other. There is no infallible statistical technique that can completely avoid ambiguity and misinterpretation, at least, not as far as we know. However, CIs offer much more information to the reader, and there is good evidence that they offer less scope for misinterpretation.

References

1. Fidler F, Cumming G, Burgman M, Thompson N. Statistical reform in medicine, psychology and ecology. *J Socio-Economics*. 2004;33:615-630. C.
2. Cumming G, Fidler F, Leonard M, et al. Statistical reform in psychology: is anything changing? *Psych Science*. 2007; 18:230.
3. Cohen J. The statistical power of abnormal-social psychological research: a review. *J Abnorm Soc Psychol*. 1962;65: 145-153.
4. Rossi JS. Statistical power of psychological research: what have we gained in 20 years. *J Consult Clin Psychol*. 1990;58: 646-656.
5. Sedlmeier P, Gigerenzer G. Do studies of statistical power have an effect on the power of studies? *Psychol Bull*. 1989; 105:309-315.
6. Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol Methods*. 2004;9:147-163.
7. Fidler F. From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology. Melbourne, Australia: University of Melbourne; 2006.
8. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med*. 1988; 321:129-135.
9. Kline RB. Beyond significance testing: reforming data analysis methods in behavioral research. *Appl Psychol Meas*. 2006;30:452-455.
10. American Psychological Association. Publication manual of the American Psychological Association. 6th ed. Washington (DC): Author; 2009.
11. Greenland S. Meta-analysis. In: Rothman KJ, Greenland S, editors. *Modern epidemiology*. 2nd ed. Philadelphia: Lippincott-Raven; 1998.
12. Kazdin AE. The meanings and measurement of clinical significance. *J Consult Clin Psychol*. 1999;67:332-339.
13. Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *J Consult Clin Psychol*. 1999;67:300-307.
14. Oakes M. Statistical inference: a commentary for the social and behavioural sciences. Chichester, UK: J. Wiley & Sons, Inc.; 1986.
15. Haller H, Krauss S. Misinterpretations of significance: a problem students share with their teachers. *Methods Psychol Res*. 2002;7:1-20.
16. Kalinowski P, Fidler F, Cumming G. Overcoming the inverse probability fallacy: a comparison of two teaching interventions. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2008;4: 152-158.
17. Cumming G. Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci*. 2008;3:286-300.
18. Cumming G, Finch S. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*. 2005;60:170-180.
19. Cumming G, Williams J, Fidler F. Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*. 2004;3: 299-311.
20. Belia S, Fidler F, Williams J, et al. Researchers misunderstand confidence intervals and standard error bars. *Psychol Methods*. 2005;10:389-396.