Teaching Confidence Intervals: Problems and Potential Solutions

Fiona Fidler
La Trobe University, School of Psychological Science,
Melbourne, Australia 3086
f.fidler@latrobe.edu.au

Geoff Cumming
La Trobe University, School of Psychological Science,
Melbourne, Australia 3086
g.cumming@latrobe.edu.au

1. Introduction

There are benefits of teaching inference via confidence intervals (CIs), rather than null hypothesis significance testing (NHST). However, CIs are not without misconceptions. First, we provide empirical evidence that CI presentations of data can help alleviate some typical misinterpretations of results, leading to more accurate conclusions and more justified decisions. Second, we demonstrate that CIs are also prone to particular types of misconceptions. Finally, we present interactive figures and simulations that, when used with guidelines for CI interpretation, should lead to more insightful interpretations of research results and fewer misconceptions.

We are studying CIs because they are the focus of reformers who are keen to reduce emphasis on NHST. Difficulties in teaching NHST are well known, and misconceptions about *p* values are widespread and severe (Haller & Krauss, 2002; Oakes, 1986; Tversky & Kahneman, 1971). Furthermore, Falk and Greenbaum (1995) demonstrated that even instructing students directly in probable pitfalls of NHST did little to improve students' interpretations of research results. CIs, as well as having many other benefits, may alleviate misconceptions associated with statistical significance. In this paper, we provide evidence for this claim, and also point out what to be wary of in teaching CIs.

2. CIs help make sense of statistical non-significance

We asked final year undergraduates and postgraduate environmental science students to interpret results presented first as *p* values and then as CIs (or vice versa). The scenarios students were asked to interpret consisted of statistically non-significant results, produced by a low powered study, but with a non-trivial (i.e. potentially ecologically important) effect size. Statistical power calculations were reported, and the scenario indicated the range of biologically important effects. In all, reporting in NHST scenarios was more detailed than is usually found in journals in ecology and conservation biology (Anderson, Burnham, & Thompson, 2000), or indeed psychology and education. Students were asked to indicate whether results provided strong or moderate support for the null hypothesis (responses of this type were considered misconceptions), strong or moderate support against the null hypothesis or whether the evidence was equivocal (all of which were considered reasonable responses).

We found that, when given *p* values, 44% (24 of 55) of students misinterpreted results as evidence for the null hypothesis—despite the important effect size and poor statistical power. Less than half as many (18%, 10 of 55) students made the same misinterpretation when results were presented with CIs. There was also evidence of a learning effect: Students who saw the CI scenario first gave the correct answer on the *p* value presentation more often

than students who saw *p* values first. The average shift in improved *p* value interpretation, after first interpreting the CI, was 1.67 points (away from responses considered misconceptions) on a 5-point Likert scale. There was no corresponding beneficial transfer of seeing the *p* values before CIs. In addition, 65% (33 of 51) of students separately indicated that NHST was either likely or very likely to mislead.

These results suggest substantial benefits of switching from a *p* value to CI approach in teaching statistical decision making. Of course, CIs have other benefits beyond decision making; for example, they may facilitate meta-analytic thinking (Cumming & Finch, 2001).

In another recent study Coulson, Fidler and Cumming (2005) sent email surveys to large numbers of authors of articles published in psychology, behavioural neuroscience or medical journals. We presented brief details of the results of two fictitious studies in an NHST format, or a CI format, and asked for some judgments about how the two studies should be interpreted. In one study the mean was statistically significantly greater than zero ($p<.05$) and in the other study it was not but, as shown in Figure 1, the two 95% CIs very largely overlapped. Respondents who saw an NHST format (one study significant, the other not) tended to see disagreement between the two studies, as we expected. Respondents who saw the CI format seemed to split into those who interpreted the CI figure in significance terms—they merely noted whether or not the CI included zero—and those who interpreted the intervals. The former group tended to see the studies as disagreeing, whereas the latter tended to see the two studies as giving consistent results. The latter interpretation is certainly more justifiable: There is no sign of a meaningful difference between the two means and a direct comparison of the two would give a large *p* value. We concluded, therefore, that CIs can potentially assist researchers and readers to avoid one of the basic problems with NHST, but will of course do so only if they are interpreted as intervals—in which case the large percentage overlap signals clearly the correct conclusion of agreement between the two studies. If CIs are merely interpreted in terms of NHST—whether or not the CI includes zero—it is hardly surprising that NHST problems persist.

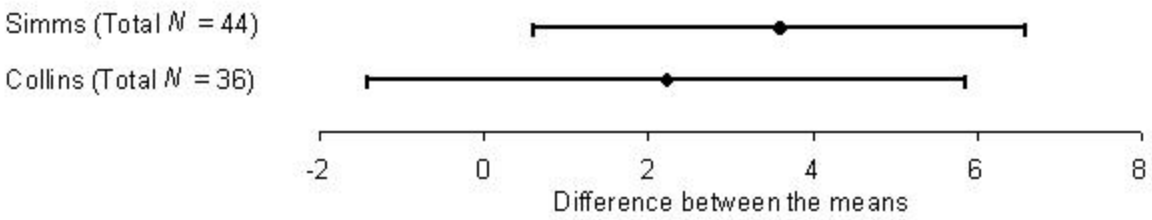


Figure 1. The graphical CI format used by Coulson et al (2005); fictitious data. Means and 95% CIs are shown for two studies. In the upper case the mean is statistically significantly greater than zero (p<.05), and in the second case it is not, but the substantial overlap of the two intervals indicate that the two studies agree.

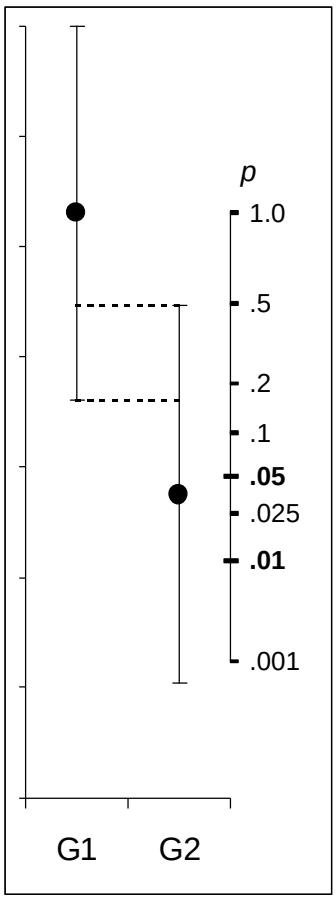
3. Beware CI misconceptions!

However, it is important to remember that CIs may have their own misconceptions. We report three studies examining how CIs are understood. In a large survey of 180 undergraduate psychology students, students displayed misconceptions over both the definition of a CI, and how aspects of a CI relate to each other. First, there is a widespread misconception that a CI is a descriptive statistic only—that is, students fail to realise the inferential nature of CIs. For example, 38% (68 of 180) believed the CI provided "plausible values for the sample mean". A further 19% (34 of 180) thought it was the "range", or "truncated range of individual scores". (These are minimum percentage values for these

misconceptions; they were much higher with other question formats.) Second, students hold a variety of misconceptions about how aspects of a CI relate to each other. For example, 20% (36 of 180) thought CI width would increase with increases in sample size; 29% (52 of 180) thought CI width was unaffected by sample size and 36% (64 of 180) were unsure if there was a relationship. Only 16% (28 of 180) could correctly answer this basic question—that, all other things being equal, CI width decreases with increases in sample size—after a full year long course in introductory statistics that even included quite a strong focus on CIs! These misconceptions were also often spontaneously generated by students in a separate study (n=95) asking for open-ended interpretations of scenario results.

Results such as those above are especially galling for the teachers who have been working hard to present and explain descriptive and inferential statistical concepts correctly. In two Internet studies we have found, however, that CI problems are not confined to students: Even world-leading researchers have a range of serious misconceptions about CIs. Cumming, Williams and Fidler (2004) sent emails to authors of articles in international journals inviting them to visit a website, where they saw a figure of a mean with its 95% CI and were asked to indicate where they felt means were likely to fall if the experiment were replicated a number of times. We found that a clear majority of respondents severely underestimated the extent to which replication means are likely to differ from the original mean. They held the *Confidence Level Misconception* (CLM), which is the erroneous belief that a *C*% CI will capture about *C*% of replication means—that the probability is about .95 that a replication mean will fall within an original 95% CI. In fact (Estes, 1997) on average only about 83% of replication means fall within an original 95% CI.

In a second Internet study, Belia, Fidler and Cumming (2005) invited researchers to a website where they saw Figure 2 (but without the dashed horizontals, or the *p* axis). They were asked to click to move the G2 mean until they judged the two means just significantly different, by independent *t* test, two-tail, α=.05. (The configuration in Figure 2 is the correct answer: The *p* value is about .05.) Responses were widely spread and few were even close to accurate. Fully 34% of respondents adjusted the mean so that the CIs lined up, end to end, with zero overlap. It is a common (but erroneous) belief in medicine (Schenker & Gentleman, 2001) that 95% CIs on independent means are just significantly different (p<.05) when the two intervals just touch end to end.



One group of respondents saw the same figure, but with the two means labelled Pretest and Posttest, and described as scores for a single group of participants. With paired data the task is, of course, not possible, because it is the CI on the *differences* that matters, not the two CIs on the separate means; a paired, not an independent *t* test is needed. A large majority of respondents in this condition overlooked the labelling of a repeated measure and responded as if the two means were independent.

We thus identified three major CI misconceptions that are held by many world-leading researchers in psychology, behavioural neuroscience and medicine. Very many of them do not appreciate how 95% CIs on two independent means give information about how the two means relate; a third believe, incorrectly, that touching end-to-end signals .05 significance; and a large majority do not appreciate the crucial role of experimental design—the importance of a repeated measure.

Figure 2. Means of two independent groups, with 95% CIs. In the Belia et al. (2005) study, respondents adjusted the G2 mean until they judged the two means to be just significantly different, a=.05. They did not see the dotted horizontals, or the p axis, which relate to our rule of eye for interpreting two independent CIs (see text).

4. Teaching CIs with strategies that overcome misconceptions

It is important and urgent to reduce emphasis on NHST, and increase use of CIs, among other techniques. Research on (mis)conception should guide development of teaching strategies for correct understanding and use of CIs. Cumming and Finch (2005) is written for the very broad audience of *American Psychologist* and offers *rules of eye*, which are intended to be easily grasped and memorable guidelines for interpretation of figures with CIs. Our first trial of these rules with students in an introductory statistics course for psychology does not mention NHST and *p* values—these come in the following semester—and bases all inference on figures showing means and 95% CIs. We judged this to be highly promising, and are planning to evaluate students' understanding, and ability to avoid CI misconceptions.

As an example, one of our rules of eye states that for two independent means, when the 95% CIs overlap by about half the average margin of error, the *p* value for the comparison of the means is about .05. Further, when the overlap is zero, the *p* value is about .01. The dotted horizontals in Figure 2 are intended to assist easy estimation of overlap. The *p* scale gives the *p* value as a function of the position of the G2 mean; it shows that our rule is a little conservative. In our course we used a modification of the rules that did not mention *p* values. In the talk we will also demonstrate some of our ESCI (Exploratory Software for Confidence Intervals, www.latrobe.edu.au/psy/esci) simulations designed to assist teaching of CIs.

REFERENCES

Anderson D.R., Burnham K.P., Thompson W.L., 2000. Null hypothesis testing: Problems, prevalence and an alternative. Journal of Wildlife Management 64, 912-923.

Belia, S., Fidler, F., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. A revision was invited by Psychological Methods, and is currently with journal reviewers.

Coulson, M., Fidler, F., & Cumming, G. (2005). Are confidence intervals better than null hypothesis significance testing? Evidence to guide statistical reform. In preparation.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 530-572.

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

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

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. Psychonomic Bulletin & Review, 4, 330-341.

Falk, R. & Greenbaum, W. (1995). Significance tests die hard. *Theory & Psychology*, *5*, 75-98.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? Methods of Psychological Research Online, 7(1). Online: www.mpr-online.de

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioural sciences. Chichester: Wiley.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*, 182-186.

Tversky, A., & Kahneman, D. (1971). Belief in the Law of Small Numbers. Psychological Bulletin, 92,105-110.