

Running Head: REPLY TO ROUDER & MOREY

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2005). Confidence intervals, still much to learn: Reply to Rouder & Morey. *Psychological Science*, 16, 494-495.

© American Psychological Society. This article may not exactly replicate the final version published in the journal. It is not the copy of record. The definitive version is available at www.blackwell-synergy.com.

Still Much to Learn About Confidence Intervals
Reply to Rouder & Morey (2005)

Fiona Fidler^{1,2}, Neil Thomason², Geoff Cumming¹, Sue Finch¹ and Joanna Leeman¹

1 = La Trobe University, Melbourne, Australia

2 = The University of Melbourne, Melbourne, Australia

Author contact:

Fiona Fidler

Department of History and Philosophy of Science

University of Melbourne, Victoria, Australia 3010

Phone: +61 3 8344 4405 Fax: + 61 3 8344 7959

Email: fidlerfm@unimelb.edu.au

We believe CIs, rather than p values, should often provide the major justification for conclusions drawn from data. Therefore CIs should be reported, and also interpreted. Rouder and Morey (2005) distinguish ‘arelational’ CIs (e.g., CIs around single sample means) and ‘relational’ CIs (e.g., CIs around mean differences or standardized effect sizes). They argue that the former are not suitable for inference and that researchers are justified in not interpreting such intervals. Yet, the purpose of research using samples is almost always to make inferences to populations. CIs—including ‘arelational’ CIs—are, by design, inferential statistics and can legitimately serve to justify inferential conclusions.

Rouder and Morey argue that, rather than using arelational CIs for inference, authors should exploit the fact that they “provide a rough guide to the variability in data, a coarse view of the replicability of patterns, and a quick check of the heterogeneity of variance” (p. 1). We believe there are problems with these three suggestions.

First, variability in data is represented directly by descriptive statistics, such as the standard deviation (SD). A CI, by contrast, is often based on a standard error (SE) and influenced by sample size. Similar levels of variability will give CIs of very different widths, depending on group size, so CIs should not be relied on to give even a rough guide to variability in data.

Second, CIs do give information about replicability, but Cumming, Williams and Fidler (2004) reported that a majority of researchers, seeing a CI, markedly underestimate the true extent of variability over replications. Further, Maxwell (2004, p. 157) pointed out that, in many realistic research situations, the pattern of results shown by CIs is unstable over replication.

Finally, to examine heterogeneity of variance, descriptive rather than inferential statistics are again needed: SDs rather than SEs or CIs. Only if group sizes are equal will CIs give an

accurate guide. Rouder and Morey's comments reinforce the need to report SDs, but do not justify non-interpretation of CIs.

CIs are rarely reported in journals outside medicine (Kieffer, Reese, & Thompson, 2001). Even in medicine, where they have for two decades been routinely reported, they are rarely interpreted (Fidler et al, 2004). We believe this is because guidelines for and examples of good practice are lacking, and we support research to develop and evaluate better guidelines for use and interpretation of CIs.

Thompson (2002) noted "It is conceivable that some researchers may not fully understand statistical methods that they (a) rarely read in the literature and (b) infrequently use in their own work" (p. 26). For example, it is widely believed (Belia, Fidler, Williams, & Cumming, 2005; Schenker & Gentleman, 2001) that two 95% CIs having zero overlap—just touching end to end—is equivalent to statistical significance with $p=.05$. In fact 95% CIs on two independent means that overlap by about one quarter of the total length of one interval correspond to p about .05 (Cumming & Finch, 2005; Saville, 2003; Wolfe & Hanley, 2002).

Rouder and Morey argue that: "Arelational confidence intervals... do not reflect between-group information and cannot be used for direct comparison" (p. 1). This is true for repeated measure designs, where CIs on separate cell means do not provide the relevant information for a comparison, but it does *not* hold for independent groups. For two independent groups the difference between the means has a p value of about .05 when the separate 95% CIs overlap by about 25% of the length of either interval, and a p value of about .01 when the two intervals just touch end-to-end (Cumming & Finch, 2005, who discuss the breadth of applicability of these rules). The terms 'arelational' and 'relational' might be useful in describing the type of CIs reported. However, such distinctions should not be used to determine the use of CIs. Of course, thought should be given to what is the most appropriate CI for the situation; as for statistical tests (Wilkinson et al., 1999). Depending on the effect of primary interest, CIs may be placed around single cell means, mean differences, interaction effects, variance accounted-for measures and so on.

Rouder and Morey rightly point out further work is required. There are no guidelines, for many situations, about how CIs should be calculated, presented and used to interpret effects. There are two main issues: Which effect (e.g., cell mean, difference, contrast, or interaction) should be plotted as a point estimate? Then, how should the relevant interval estimate be displayed?

Gardner and Altman (1986) advised that "the major contrasts of a study should be shown directly, rather than only vaguely in terms of the separate means" (p. 748). They advocate plotting mean differences rather than separate means, as in Rouder and Morey's plotting in Figure 1, right panel, of point estimates of effects of interest (except, perhaps, for the effect sizes being standardized, but that is another issue).

Estes (1997) recommended that a CI should only be attached to a point estimate if it is calculated from the variability of the data on which that point estimate is based. Following Estes' policy would mean that Error Bars C should not be attached to the cell means in Figure 1, left panel.

Loftus and Masson (1994) and Mason and Loftus (2003) recommend attaching CIs calculated for a repeated measure variable to all of the cell means involved; this violates Estes' (1997) recommendation, but is consistent with Rouder and Morey's display of Error Bar C in Figure 1.

These remain issues for debate. The advantages of CIs make the debate worthwhile. First, looking at CIs across studies should facilitate a meta-analytic approach, leading eventually

to a true parameter value, even if our original expectations were wildly wrong (Schmidt, 1996). Second, a focus on estimation should improve the way psychologists theorize, and plan and conduct empirical research.

References

- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*.
- 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*, 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.
- Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62*, 749-770.
- Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750.
- Kieffer, K.M., Reese, R.J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, *69*, 280-309.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.
- Rouder, J. N., & Morey, R. D. (2005). Relational and arelational confidence intervals: A comment on Fidler et al. (2004). *Psychological Science*, *16*, 77-79.
- Saville, D. J. (2003). Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology*, *57*, 167-175.
- 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.
- Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 24-31.
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.
- Wolfe, R., & Hanley, J. (2002). If we're so different why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal*, *166*, 65-66.

Author note

We may be contacted by email: Fiona Fidler: fidlerfm@unimelb.edu.au, Geoff Cumming: G.Cumming@latrobe.edu.au. This work was supported by the Australian Research Council.