Statistical Reform in Psychology : Is Anything Changing?

Geoff Cumming, Fiona Fidler, Martine Leonard, Pavel Kalinowski, Ashton Christiansen, Anita Kleinig, Jessica Lo, Natalie McMenamin and Sarah Wilson

Psychological Science 2007 18: 230
DOI: 10.1111/j.1467-9280.2007.01881.x

Short Report

Statistical Reform in Psychology

Is Anything Changing?

Geoff Cumming, Fiona Fidler, Martine Leonard, Pavel Kalinowski, Ashton Christiansen, Anita Kleinig, Jessica Lo, Natalie McMenamin, and Sarah Wilson

La Trobe University, Melbourne, Victoria, Australia

We investigated whether statistical practices in psychology have been changing since 1998. Early in this period, the American Psychological Association (APA) Task Force on Statistical Inference (TFSI; Wilkinson & TFSI, 1999) advocated improved statistical practices, including reporting effect sizes and confidence intervals (CIs). The APA (2001) *Publication Manual* included expanded advice on effect sizes and stated that CIs are "in general, the best reporting strategy" (p. 22). Any changes since 1998 may, of course, have had many causes other than those two sources of advice.

STATISTICAL PRACTICES IN LEADING JOURNALS

We examined 10 leading international psychology journals that publish mainly empirical research; Figure 1 shows the journals. For each journal, we coded 40 empirical articles published in each of three periods: the first 40 published in 1998, the most recent 40 available for 2003–2004 when we coded in mid-2004, and the most recent 40 available for 2005–2006 around April 2006. We focused on three statistical practices central to the statistical-reform debate that we could code reliably: We noted any use of null-hypothesis significance testing (NHST), CIs, and figures with error bars. In addition, we searched for any discussion of CIs or error bars, or use of these to support interpretation. (Finch et al., 2004, found that even when CIs are reported, they are rarely interpreted.) We did not code effect sizes, but have elsewhere documented increases in some effect size reporting, for example, in the *Journal of Consulting and Clinical Psychology* (*JCCP*; Fidler et al., 2005).

NHST was used in almost all articles (97.8, 97.7, and 96.9% in the three time periods, respectively). There was a substantial increase in figures with error bars (11.0, 24.7, and 37.8%; see Fig. 1, left panel; note that the CI around each mean difference does not include the mean of the previous time period, which indicates a statistically significant change from the previous time period, $p < .05$, two-tailed; or $p_{rep} > .92$). The error bars were sometimes CIs (12.7, 9.5, and 13.8% of articles with figures with error bars), but more often represented standard errors (58.7, 43.1, and 46.7%) or were not labeled (33.0, 28.8, and 33.8%)—a serious omission (Cumming & Finch, 2005). The *Journal of Experimental Psychology: General* (*JEPG*; 12.5, 62.5, and 75%) showed the largest increase in figures with error bars, and every journal except *Child Development*, the *Journal of Abnormal Psychology*, and *JCCP* showed an increase from 1998 to 2005–2006 that was larger than the lower arm of the CI around the difference between those two times (examples for some journals are shown by the light error bars in Fig. 1, left panel).

CIs were rarely reported, but their use increased somewhat (3.7, 9.2, and 10.6%; Fig. 1, right panel). The mean increase from 1998 to 2003–2004 was larger than the lower arm of the CI around the difference. The *Journal of Abnormal Psychology*, *JCCP*, and *JEPG* showed the largest increases, and they and *Child Development* and the *Journal of Abnormal Child Psychology* showed increases larger than the lower arm of the CI around the difference between 1998 and 2005–2006. Most CIs were reported in the text or a table (78.7, 64.2, and 71.5%), rather than as error bars in figures. Only 24.1% of articles with CIs had any interpretation of a CI, and such interpretation was often in terms of NHST, rather than width or precision (Cumming & Finch, 2005). Only 3 of the 168 articles with standard error bars in a figure included any interpretation of error bars.

We examined editorials, Web sites, and instructions to authors for the 10 journals. Most offered little statistical advice, and there was little change since 1997. Exceptions were the Web site for *JEPG*, which since June 2004 has provided extensive statistical advice (http://web.uvic.ca/jepgen/tips.htm); editorials in *JCCP* (Fidler et al., 2005; La Greca, 2005); and the instructions to authors for *Psychological Science*, which since March 2004 have required effect sizes, and since July 2005 have encouraged authors to use p_{rep} (Killeen, 2005), not NHST.

OPINIONS OF AUTHORS AND EDITORS

We e-mailed questions to contact authors of our 2003–2004 articles, using different return addresses so we could separate

Address correspondence to Geoff Cumming, School of Psychological Science, La Trobe University, Victoria, Australia 3086, e-mail: g.cumming@latrobe.edu.au.

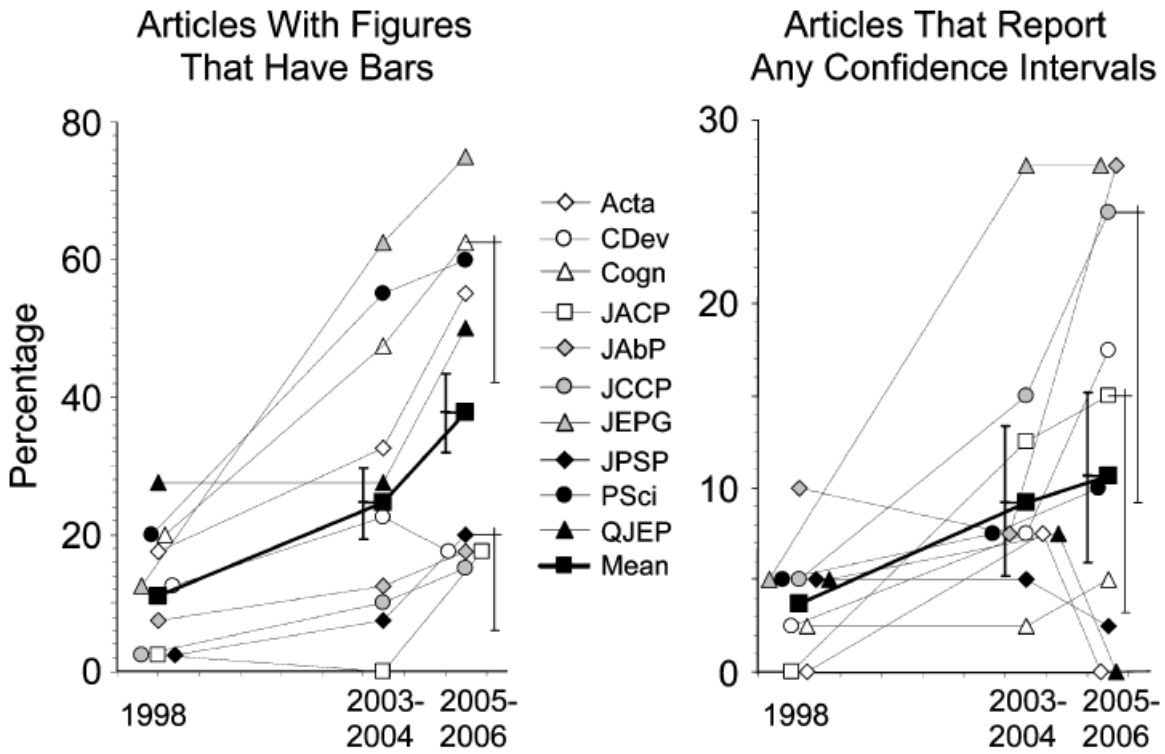

Fig. 1. Percentages of articles that included a figure with error bars (left panel) or included a confidence interval (CI; right panel) in text, in a table, or as error bars in a figure, for 10 journals in each of three time periods. Note the differing vertical scales in the two panels. Forty articles were examined for each journal at each time, so 2.5% represents one article; readers should be cautious of overinterpreting fluctuations for individual journals. Points have been adjusted horizontally where necessary to avoid overlay. Filled squares and heavy lines show means. Heavy error bars are 95% CIs on the difference in percentage between adjacent times. Light error bars are examples of lower-half 95% CIs on the difference in percentage for individual journals between 2005–2006 and 1998. Acta = *Acta Psychologica*, CDev = *Child Development*, Cogn = *Cognition*, JACP = *Journal of Abnormal Child Psychology*, JAbP = *Journal of Abnormal Psychology*, JCCP = *Journal of Consulting and Clinical Psychology*, JEPG = *Journal of Experimental Psychology: General*, JPSP = *Journal of Personality and Social Psychology*, PSci = *Psychological Science*, QJEP = *Quarterly Journal of Experimental Psychology* (Section A only, before 2006).

responses from authors who had and had not used CIs or error bars in their article. We received 102 responses, a 29% response rate. Respondents stated that they had moderate awareness of the statistical-reform debate, but less awareness of the statistical recommendations of Wilkinson and TFSI (1999) and the APA (2001) manual. Only 30% felt current statistical practices are satisfactory; 55% thought CIs should be used more widely, and 75% thought CIs provide additional useful information over NHST, but 55% felt NHST has served psychology well. The most common additional comment expressed support for effect sizes. Of course, our respondents were self-selecting. There was no sign of any differences of views between authors who had and had not used CIs or error bars in their articles.

In August 2005, we e-mailed questions to editors, and received responses from nine. They made many insightful comments and expressed a range of statistical concerns. Five editors emphasized the need for effect size reporting, and four advocated wider use of error bars, but freedom to match statistical techniques to the situation was generally preferred over any strict statistical requirements. Beyond mention of the interesting *Psychological Science* experiment with p_{rep} (Cumming, 2005), there were few indications of intentions to seek changes to statistical practices.

CHANGE, BUT LITTLE REFORM YET

At least in these 10 journals, NHST continues to dominate overwhelmingly. CI reporting is increasing but still low, and CIs are seldom used for interpretation. Figures with error bars are now common, but bars usually indicate standard errors, not the recommended CIs (Cumming & Finch, 2005). It is a persisting problem that many error bars are not identified. Even when included in figures, standard error bars are almost never used for data interpretation.

Some comments from authors and editors indicated disquiet about current practices and some readiness to contemplate change, although statistical reform is not currently an editorial priority. Some authors identified journals' expectations as the obstacle; a few editors mentioned authors' resistance. NHST is undoubtedly deeply ingrained in psychologists' thinking.

It is important and urgent that psychology change its emphasis from the dichotomous decision making of NHST to estimation of

effect size (Gigerenzer, 1998; Meehl, 1978). Effect sizes must always be reported—in an appropriate measure, and wherever possible with CIs—and then interpreted. To achieve this goal, researchers need further detailed guidance, examples of good practice, and editorial or institutional leadership.

Acknowledgments—We thank the authors and editors who responded to our questions, and Cathy Faulkner for comments.

REFERENCES

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

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

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

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., & Schmitt, R. (2005). Evaluating the effectiveness of editorial policy to improve statistical practice: The case of the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology*, *73*, 136–143.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of *Memory & Cognition*. *Behavior Research Methods, Instruments, & Computers*, *36*, 312–324.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*, 195–204.

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

La Greca, A.M. (2005). Editorial. *Journal of Consulting and Clinical Psychology*, *73*, 3–5.

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.

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

(RECEIVED 6/16/06; ACCEPTED 6/16/06;
FINAL MATERIALS RECEIVED 6/23/06)