

Running head: STATISTICAL REFORM IN OTHER DISCIPLINES

Fidler, F. & Cumming, G. (2007). Lessons learned from statistical reform efforts in other disciplines. *Psychology in the Schools*, 44, 441-449. DOI: 10.1002/pits.20236

© Wiley Periodicals, Inc. Journal website:

<http://www.wiley.com/WileyCDA/WileyTitle/productCd-PITS.html> This article may not exactly replicate the final version published in the journal. (It is the version just before copy editing.) It is not the copy of record.

Lessons Learned From Statistical Reform Efforts in Other Disciplines

Fiona Fidler and Geoff Cumming

La Trobe University, Melbourne, Victoria, Australia

Fiona Fidler

School of Psychological Science,

La Trobe University, Victoria, Australia 3086

Email: f.fidler@latrobe.edu.au

Compelling arguments for reform of statistical practices have been made in many disciplines, in some cases over several decades, but achieving reform has proved difficult. The authors discuss how reform has progressed—or not progressed—in psychology, medicine and ecology, and describe case studies of attempts by pioneering journal editors to change statistical practices. Lessons for those seeking reform in education include the need to recognize the importance of journal editors, and of provision of articles, books and software that give practical guidance to researchers wishing to use the recommended statistical techniques. Research is required on recommended techniques so that statistical practice can become evidence-based. Also, improvement in statistical practice should be encouraged along with improvement in the way a discipline theorizes.

The statistical reform ‘debate’ has been, very largely, the sound of one hand clapping. Leading scholars have published cogent arguments, with evidence, that null hypothesis significance testing (NHST) is widely misunderstood and misused, and causes serious damage to research progress. It is hardly a debate, however, because there have been few defenses of NHST, and those brave writers who have advocated use of NHST usually concede many of the problems, and can only try to argue that, even so, there may be some value of NHST in certain particular circumstances. Moreover, the devastating critiques of NHST and the way it is used have been repeated decade by decade (Finch, Thomason, & Cumming, 2002), and in the journals of numerous disciplines (Altman, 2004; Anderson, Burnham, & Thompson, 2000). Such criticisms can be found in a wide range of disciplines: In addition to psychology, medicine and ecology, they can be found with ease in: sociology (e.g., Morrison & Henkel, 1970); education (e.g., Thompson, 1996); criminology (e.g., Weisburd, Lum & Yang, 2003); economics (e.g., Zilak & McCloskey, 2004); marketing (e.g., Sawyer & Peter, 1983); chemistry (e.g., Harris, 1993); nursing (e.g., Glaser, 1996) and, notably, statistics itself (e.g., Royall, 1986). The case for reform is compelling, and we believe reform is both necessary and urgent (Cumming & Finch, 2005; Thompson, 2006b).

Even so, NHST remains overwhelming the way researchers draw conclusions from data in many disciplines, including education and across the social and behavioral sciences. How can practice be at such stark variance from expert recommendations?! There are important sociological reasons why everyone, from students to teachers to researchers to journal editors, finds it very hard to change long-established practices. The textbooks, statistics curricula and available software are all centered on NHST. Virtually every journal article exemplifies NHST practice. Researchers write what they know journal editors will publish, and editors wish to attract the top scholars, and usually feel they should allow those scholars to choose whatever techniques they think best to present their work.

An additional reason for the persistence of NHST is that the misconceptions and cognitive fallacies associated with NHST are so strong and intuitively appealing (Schmidt & Hunter, 1996). Many of these center on misunderstanding of the p value. Cohen (1994) explained the inverse probability fallacy, which is a confusion between the p value—which is $P(D|H)$, the conditional probability of our data (or more extreme), given our null hypothesis—and what we really want to know, which is $P(H|D)$, the conditional probability that our hypothesis is true given the data. He argued that the p value “does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!” (p. 997). Oakes (1986) presented evidence on the extent of misconception, and also identified the inverse probability fallacy as a major reason why NHST is widely used, but widely misunderstood. Haller and Krauss (2002) demonstrated that this misconception has not diminished over time, and is widespread even amongst statistics teachers.

We believe there is also a deeper reason why NHST persists, and it relates to the way theorizing is conducted in a discipline. NHST is oriented towards making dichotomous decisions: The null hypothesis is either rejected, or not rejected. In a discipline in which the level of theoretical discourse primarily concerns whether or not one variable has an effect on another, it is natural to choose a statistical tool that, correspondingly, gives a dichotomous outcome. NHST does this, and can even go a little further and permit a conclusion as to the direction of the effect the first variable has on the second. Such primitive theorizing would be laughed out of court in many disciplines: Imagine a theory of gravity that merely said “there is a force between all bodies”, or at most could say “it is an attractive force”. Or consider a chemist who reports that the melting point of an important new substance “is significantly higher than zero”! Gigerenzer (1998) argued that in psychology many theories are really only “surrogates” for theories, in that they are merely simple-minded dichotomies, or mere directional statements of a relation, and that the dominance of the simplistic dichotomous decisions of NHST helps perpetuate such an impoverished approach to theorizing.

So, there are many reasons why NHST persists. A further difficulty for reform is that the alternative techniques advocated by reformers have received little cognitive examination, and so it is not clear that replacement techniques will lead to better research communication, with less misconception than NHST. For example we have published two studies of how researchers in psychology, behavioral neuroscience, and medicine think about and interpret confidence intervals (CIs) (Belia, Fidler, Williams, & Cumming, 2005; Cumming, Williams, & Fidler, 2004). We identified five misconceptions that appear to be widely held about how CIs should be interpreted in various situations. More immediately, when considering alternatives to NHST Cohen (1994) remarked “I suspect that the main reason [confidence intervals] are not reported is that they are so embarrassingly large!” (p. 1002). It appears that successful statistical reform, which results in statistical practices that are statistically justified and correctly understood by users, is a very challenging goal—even if it is essential for efficient and effective research.

A disappointing feature of the statistical reform debate, and statistical reform efforts, is that similar arguments are made, similar difficulties identified, and similar solutions proposed, in discipline after discipline, with little inter-disciplinary communication. There is a general lack of realization of what is happening in other disciplines. Reform will surely be highly laborious if every discipline must make the same mistakes and learn the same lessons for itself?

In this article we now consider statistical reform in psychology, medicine, and ecology, and seek lessons for education. We gave an earlier and more detailed discussion of these three disciplines in Fidler, Cumming, Burgman, and Thomason (2004). That article includes many more references. These three disciplines are especially interesting to consider because their reform stories are so different.

Statistical Reform in Psychology

In the late 1940s and 1950s experimental psychology developed rapidly. It was especially concerned to be objective, empirical and scientific, and this was a major reason for its rapid adoption of NHST as its dominant way to draw conclusions from data (Gigerenzer, 1987). NHST seemed to offer scientific certainty, even if this was often an illusion, especially in a discipline in which Type II error rates are high and usually unknown (Cohen, 1962). By the late 1950s, 86% of empirical articles in leading American Psychological Association (APA) journals used NHST (Hubbard & Ryan, 2000). In little more than a decade NHST had leapt to statistical dominance.

The cogent critiques in psychology started at that time (e.g., Meehl, 1954) and have continued ever since. Finch et al. (2002) recounted the decade-by-decade critiques, and also described how successive revisions of the *APA Publication Manual* recommended that statistics be reported in journal articles. The *Manual* is extraordinarily influential: Psychology students around the world follow its guidelines as they learn the ‘APA style’ in which their laboratory reports must be written. More than 1,000 journals across education and the social and behavioral sciences, and beyond, refer to the *Manual* in their advice to prospective authors. Although its purpose is primarily as a style guide, with numerous details of things such as use of italics, accepted abbreviations, and how to format references, there is also advice on how to report statistical results. Finch et al. reported that the successive editions of the *Manual* generally made few changes in recognition of the statistical reformers’ arguments, and that NHST techniques continued to dominate in the *Manual*’s statistical advice. In the 1994 edition (APA, 1994) it was recommended that effect size measures be reported and statistical power considered seriously, and in the current edition (APA, 2001) there is a recommendation that CIs “are, in general, the best reporting strategy” and are “strongly recommended” (p. 22). However there was not a single example of CI use, and no recommendation as to how CIs should be reported or interpreted, although there remain numerous recommendations about NHST. The overall message of the current manual is ‘NHST business as usual’, and many reformers found this extremely disappointing and a major reform opportunity lost (Fidler, 2002). Journal surveys have shown that even the modest reform recommendations of those two editions of the *Manual* have had little influence on practice.

One of the few attempts by a journal editor in psychology to achieve substantial statistical change was that of Geoff Loftus at the major journal *Memory & Cognition* (*M&C*). As incoming editor he explained that he regarded NHST as a poor way to draw from data the conclusions that are likely to be of most research interest (Loftus, 1993). He urged authors to report error bars—either standard error (SE) bars, or CIs—in graphs, and to rely on these to support their inferences. He urged authors to omit NHST entirely: This was a radical move! Finch et al. (2004) scrutinized articles published in *M&C* between 1990 and 2000. They found that use of error bars did indeed

increase during the Loftus years (1994-1998). It peaked at 47% of articles, but dropped after he finished as editor. Almost no articles omitted NHST entirely, and even when error bars were reported they were seldom used to justify data interpretation. Loftus achieved considerable change in what M&C authors reported, but little change in how they justified their conclusions—which remained overwhelmingly by use of NHST. Loftus found the process exhausting and frustrating, and Finch et al. could find few lasting changes resulting from his efforts. Even a determined and energetic editor could, it seemed, achieve little change in statistical practices.

In 1997 in the *Journal of Consulting and Clinical Psychology* the new editor Philip Kendall (1997) encouraged prospective authors to report clinical significance as well as statistical significance. Fidler et al. (2005) examined articles published in that journal and found little change in the discussion of clinical significance from 1996 to 2000-01. Again, editorial encouragement appeared to be able to achieve little change.

In 1996 APA set up the Task Force on Statistical Inference to consider a suggestion that NHST be banned from APA journals. The report of the Task Force (Wilkinson et al, 1999) is an excellent source of advice on experimental design and statistics. It did not recommend a ban on NHST, but it explained why effect sizes should always be reported and it mentioned CIs with approval and supported wider use of figures with error bars.

Our conclusion is that the numerous articles advocating statistical reform in psychology have to date had relatively little impact on practice. Recent small signs of potential change include the publishing of important books explaining effect size, estimation, and other recommended techniques (e.g., Grissom & Kim, 2005; Kline, 2004), and textbooks that place CIs centre stage (e.g., Smithson, 2000; Thompson, 2006a). Hill and Thompson (2004) identified 23 psychology and education journals that have an editorial policy encouraging use of effect sizes, or other preferred techniques, or warning of problems with NSHT. Even so, NHST remains the focus of almost all textbooks, statistics courses and widely-used statistical software, and is still the basis for interpretation in almost all empirical articles in psychology journals. Psychology has also provided case studies suggesting that individual editors are unlikely to achieve substantial changes in practice.

As a recent footnote to the psychology story we note the leading journal *Psychological Science* published in 2005 a major article by Killeen (2005), which argued that inference should be based not on NHST but on p_{rep} , the probability that a replication of a study would give a result in the same direction as the original study. *Psychological Science* has since published several reactions to that proposal (e.g., Cumming, 2005), and since mid-2005 has encouraged authors to report p_{rep} , and to use it rather than p values for interpretation of their data. The p_{rep} proposal is explained in this issue by Killeen himself (2006).

Statistical Reform in Medicine

In the 1950s and 1960s NHST became established as standard practice in medicine, as part of the development of the clinical trial. Therapeutic reformers championed the clinical trial, concerned that decisions made simply on the expert recommendation of individual physicians were too time consuming and too open to biases and pressure from pharmaceutical companies. The still relatively obscure NHST techniques appeared to possess the qualities they were looking for: efficiency and objectivity. NHST was adopted as part of the standard clinical trial procedure and rapidly became the primary method for assessing new drugs and justifying their use.

Soon there were critics, for example Cutler, Greenhouse, Cornfield, and Schneiderman (1966), who felt that statisticians rather than clinicians were becoming primary decision makers. During the 1970s criticism of NHST increased, and led to advocacy of CI use. Editors of leading

journals took serious note of the issue. The *New England Journal of Medicine* published editorials encouraging CI use and explaining problems of NHST (Rennie, 1978; Rothman, 1978).

Ken Rothman was an early advocate of reform. As assistant editor of the *American Journal of Public Health (AJPH)* from 1983 he advocated use of CIs, and routinely wrote to submitting authors asking that they delete all p values, or other use of NHST, from their manuscripts—or seek to publish elsewhere. In 1990 Rothman, as founding editor of *Epidemiology*, stated that “When writing for *Epidemiology*, you can enhance your prospects if you omit tests of statistical significance... we do not publish them at all” (Rothman, 1998, p.9). Fidler, Thomason, Cumming, Finch, and Leeman (2004) assessed the effectiveness of Rothman’s policies, and found in *AJPH* a striking increase in CI use from 10% before Rothman to 54% in his time, and that reliance on p values alone decreased from 63% pre-Rothman to just 5%. His impact at *Epidemiology* was even more dramatic: In 2000, 94% of articles reported CIs and none reported p values.

During the 1980s, editors of some other leading journals also took initiatives designed to warn about NHST problems and encourage use of CIs. By 1988, more than 300 medical and biomedical journals had adopted the guidelines of the International Committee of Medical Journal Editors (ICMJE), which included the statement:

When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of p values, which fail to convey important quantitative information... (ICMJE, 1988, p.260).

Important for statistical change in medicine was the publishing of expert but accessible guidance and software to help researchers adopt recommended practices. For example *BMJ* Books published *Statistics with confidence* in 1989; the second edition is Altman, Machin, Bryant, and Gardner (2000). CIs have now for 20 years been routinely reported in empirical articles in medicine, but p values are still common, and conclusions are often based on NHST rather than the reported CIs. Statistical reform may have progressed further in medicine than psychology, but statistical practice in medicine is not yet ideal.

An important development in medicine has been the rise of meta-analysis, which is the quantitative combination of results over a number of studies that have reported evidence on the same, or closely related questions. Any individual experiment should be planned in the context of a meta-analysis of previous similar studies, and then its results entered into that meta-analysis. The Cochrane Collaboration (www.cochrane.org) is the largest database of medical meta-analyses. Most of these summarize the evidence they are combining by means of forest plots, which are figures that represent the result of each experiment on a particular question as a single effect size, with a CI. Then the effect size for all the studies combined is also displayed, with its CI. The emergence of meta-analysis has thus reinforced CI use.

A further crucial development has been the widespread adoption of *evidence-based practice*. To be professional and ethical, clinical practice must be justified by research evidence, as summarized by, for example, a Cochran meta-analysis. Improved statistical practices, especially CIs and meta-analysis, play a central role in analyzing and presenting the evidence base, and thus in shaping what is now regarded as medical best practice.

Statistical Reform in Ecology

Ecology field work has particular challenges for the researcher interested in rigorous experimental design. Populations of plants or animals are often small, and individuals may be hard to find; randomization may be impractical. Such difficulties may help explain why NHST

only became established gradually in ecology through the late 1950s and 1960s, although it did eventually come to dominate as the way conclusions are drawn from data (Fidler, Burgman, Cumming, Buttrose, & Thomason, in press). However NHST presents particular problems in ecology because low and unknown statistical power can easily lead to Type II errors with devastating consequences. For example, small threatened populations or fragile environments are unlikely to be identified for ecological rescue if a low power experiment fails to find statistically significant risk or damage. The consequences of such errors can be irreversible.

Debate about NHST in ecology emerged only in the 1980s, and mainly concerned statistical power, with little concerted advocacy of CIs or other preferable techniques. While understanding and reporting power can be valuable, it is a concept that has meaning only in the context of NHST, so a focus on power is likely to reinforce use of NHST at the expense of estimation. In addition, discussion in ecology about power became somewhat side-tracked into debate about the value of post hoc power, which is based on the observed effect size and which is of little or no value (Hoenig & Heisey, 2001).

A promising development, however, is the increasing use in ecology of modeling, for example the development of various models of how population size or composition changes, and then testing of the fit and sensitivity of those models. Models that explain data well can give insight into underlying processes, and may permit prediction of likely future developments in a population. Models do not have to be very good to be much more useful than a mere dichotomous decision based on NHST—especially when that is likely to be a Type II error! Model testing and selection has encouraged the use of information theoretic and Bayesian methods, and use of such methods is likely to expand as they become better-known to researchers, and as helpful books and software become available. They hold great promise for ecological research, and it is notable that statistical advance is occurring along with advances in how researchers in ecology conduct their theorizing. There is very recent evidence of change in statistical practices in at least some areas of ecology. In 2001 and 2002, 92% of sampled articles in *Conservation Biology* and *Biological Conservation* reported at least one *p* value. In 2005, this figure had dropped to 78% (Fidler, Burgman, Cumming, Buttrose, & Thomason, in press).

Conclusions About Reform in Other Disciplines

Change is difficult

We have not found any case of change being achieved quickly or easily, and we have seen cases (e.g., Loftus at *M&C*) where even determined efforts achieved only change that was limited and did not endure.

Requirements may be more effective than recommendations

Rothman at *AJPH* and *Epidemiology* achieved change that was greater and more enduring than that achieved by Loftus. Rothman's *requirements* of authors can be contrasted with the strong *encouragement* of Loftus. However Rothman was operating in a context in which medicine was already starting to provide resources—articles, books and software—to assist authors use the new techniques, and other editors were starting to make policy encouraging the changes. Loftus, by contrast, had in psychology little such contextual support for his efforts.

Editors are influential, but it's best if they work together

Sole editors may be able to achieve little, but once a critical mass of leading medical editors became convinced, the policy statement by the ICMJE quickly spread the new requirements to hundreds of journals.

Institutions may play a role?

Every discipline has its particular institutions. The effectiveness of the ICMJE suggests that institutions can sometimes play a useful role. In psychology the *Manual* has a central place, and it is influential across many other disciplines as well. If the APA or its *Manual* were to take a strong policy position, that might help achieve reform, but the weak statements in the current *Manual* have so far had little effect.

Books, tools and guidance matter

It seems obvious that authors will not do what they do not understand, and cannot readily carry out. Changes in statistical practices in medicine were accompanied by explanatory journal articles, books, and software all designed to be accessible to medical researchers and to support the practical calculations required by the recommended practices. Statistics textbooks and courses also need to present the desired techniques. We suspect that provision of all these resources is crucially important for achieving widespread reform; without easy-to-follow guidelines and examples of good practice, researchers are unlikely to use a technique, no matter how strong the exhortation.

Action is needed on many fronts

It is safe to conclude that a reform advocate should proceed on many fronts: Make the case for reform, provide resources and guidance so it is easy to do what is recommended, and try to persuade editors and the discipline's influential institutions to take a strong stand. It may be a long and challenging road, but the goal is vitally important, so persist!

Statistical practice is integrated into a discipline's thinking

Choice of statistical practices is integrated into how a discipline does its thinking, theorizing and researching. NHST was quickly adopted by psychology and medicine in years in which those disciplines were seeking to be objective and scientific, and wanted statistical techniques to justify decisions. To some extent psychology still theorizes in terms of simple relations between variables, so NHST is still seen by many to be adequate. In medicine there was increasing concern with clinical significance—which focuses on the size of an effect, not merely whether it was positive or negative—and so the dichotomous decisions of NHST were inadequate and CIs needed. In ecology, the growth of sophisticated model building and model selection demands sophisticated statistical techniques to assess the fit of models to data. In general, changes to statistical practice may encourage, or be encouraged by, changes in theorizing, and it may be most useful for statistical reformers—as well as those interested in promoting improvements in how a discipline theorizes—to consider the two as operating synergistically, and to work for advances on both fronts.

Statistical practice should be evidence-based

Evidence-based practice is an ideal that is spreading from medicine to many other fields, including clinical psychology, business, and education. We believe statistical practice should also

be evidence-based: Reformers should be able to cite evidence that the techniques they advocate as better than NHST will indeed reduce misconception and improve research communication. To date there is little such evidence; research is required. However, NHST problems are so severe, and the case for CIs and other techniques so strong, that reform should not be delayed, but research is needed to improve further the statistical techniques recommended as best practice. An overarching goal of this work must be to develop further links between two parallel literatures: the empirical literature on statistical reasoning and decision making (e.g., Sedlmeier, 1999) and the rhetorical literature of statistical reform.

Lessons for Education

Some fields in education use path analysis and other sophisticated statistical techniques to develop and test detailed models. Measurement and testing is another educational field in which advanced statistical techniques have been developed and widely applied. However, in much of education, as still in much of psychology, many questions are still examined in a simplistic dichotomous way: Does X improve Y, or not? We believe all our conclusions above can be taken as lessons for education, but we would like to emphasize the last two. First, don't just advocate a switch from NHST to, for example, CIs, valuable though that may be. Work also to enrich the questions being asked: By how much does X improve Y? And with what precision does our experiment give an answer to that question? Such reconceptualization naturally calls for use of CIs. An advance in thinking and an advance in statistical technique: A double benefit! Second, encourage research on statistical practices in education, so that educational research and practice can be served by statistical practice that is evidence-based, and thus making as strong as possible a contribution to progress in the discipline.

References

- Altman, M. (2004). Introduction. *Journal of Socio-Economics*, 33, 523-525
- Altman, D.G., Machin, D., Bryant, T.N. & Gardner, M.J. (Eds.). (2000). *Statistics with Confidence* (2nd ed.). London: BMJ Books.
- American Psychological Association (1994). *Publication Manual of the American Psychological Association* (4th ed.). Washington, DC: Author.
- American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Anderson, D., Burnham, K., & Thompson, W. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64, 912-923.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10, 389-396.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- 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.
- Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, 3, 299-311.
- Cutler, S.J., Greenhouse, S.W., Cornfield, J., & Schneiderman, M.A. (1966). The role of hypothesis testing in clinical trials. *Journal of Chronic Diseases*, 19, 857-882.

- 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.
- Fidler, F., Burgman, M., Cumming, G., Buttrose, R. & Thomason, N. (in press). Impact of criticisms of null hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology*.
- Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics*, *33*, 615-630.
- 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.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals but they can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119-126.
- 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.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, *12*, 825-853.
- Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Kruger, J. Lorraine & G. Gigerenzer (Eds.). *The Probabilistic Revolution, Vol. 2: Ideas in the sciences* (pp. 11-33.). Cambridge, MA, USA: The MIT Press.
- Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*, 195-204.
- Glaser, D.N. (1996). The need for a moratorium on significance testing. *Journal of cardiovascular nursing*, *11*, viii-ix.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research. A broad practical approach*. Mahwah, NJ: Erlbaum.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*, 1-20.
- Harris, E.K. (1993). On *p* values and confidence intervals (why can't we *p* with more confidence?). *Clinical chemistry*, *39*, 927-928.
- Hill, C.R., & Thompson, B. (2004). Computing and interpreting effect sizes. In J.C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 19, pp. 175-196). New York: Kluwer.
- Hoening, J.M., & Heisey, D.M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19-24.
- Hubbard, R. & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement*, *60*, 661-681.
- International Committee of Medical Journal Editors. (1988). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine*, *108*, 258-265.
- Kendall, P.C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *65*, 3-5.
- Killeen, P. R. (2005). An alternative to null hypothesis significance tests. *Psychological Science*, *16*, 345-353.
- Killeen, P. R. (2006). Why not replace confidence intervals with the probability of replicating an effect? *Psychology in the Schools*, *xxx*, yyy-zzz.

- Kline, R. B. (2004) *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.
- Loftus, G. (1993). Editorial comment. *Memory & Cognition*, 21, 1-3.
- Meehl, P.E. (1954). Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.
- Morrison, D.E., & Henkel, R.E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.
- Oakes, M. W. (1986). Statistical inference: a commentary for the social and behavioural sciences. Chichester, U.K.: Wiley.
- Rennie, D. (1978). Vive la difference ($p < 0.05$). *New England Journal of Medicine*, 299, 828-829.
- Rothman, K.J. (1978). A show of confidence. *New England Journal of Medicine*, 299, 1362-1363.
- Rothman, K.J. (1998). Writing for Epidemiology. *Epidemiology*, 9, 333-337.
- Royall, R. (1986). The effect of sample size on the meaning of significance tests. *American Statistician*, 40, 313-315.
- Sawyer, A.G., & Peter, J.P. (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research*, 20, 122-133.
- Schmidt, F., & Hunter, J. (1996). Eight common but false objections to the discontinuation of significance testing in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.
- Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Erlbaum.
- Smithson, M. (2000). *Statistics with confidence*. London: Sage.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher*, 25, 26-30.
- Thompson, B. (2006a). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.
- Thompson, B. (2006b). The justification for the use of confidence intervals and effect sizes to report school psychological research. *Psychology in the Schools*, xxx, yyy-zzz
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
- Weisburd, D., & Lum, C.M., & Yang, S.M. (2003). When can we conclude that treatments or programs "don't work"? *The Annals of the American Academy of Political and Social Science*, 587, 31-48.
- Ziliak, S. & McCloskey, D. (2004). Size matters: The standard error of regressions in the *American Economic Review*, *Journal of Socioeconomics*, 33, 665-675.

Author Note

Fiona Fidler and Geoff Cumming, School of Psychological Science, La Trobe University. Correspondence about this article may be addressed to Fiona Fidler or Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Australia 3086. Email: f.fidler@latrobe.edu.au or g.cumming@latrobe.edu.au.