SHOULD PSYCHOLOGY ABANDON *p* VALUES AND TEACH CIs INSTEAD? EVIDENCE-BASED REFORMS IN STATISICS EDUCATION

Fiona Fidler
La Trobe University, Australia
f.fidler@latrobe.edu.au

Several editorial and institutional interventions in psychology have aimed to improve statistical reporting in journals. These efforts have sought to de-emphasise statistical significance and encourage alternative analyses, especially effect sizes and confidence intervals (CIs), but the interventions to date have had short-lived and superficial impact—if any impact at all. I review some of these interventions in psychology and discuss possible reasons for lack of success. I give an inter-disciplinary context by discussing reform efforts in medicine—in which useful reform has already been achieved—and ecology. I then identify statistics education as the next major challenge for reformers, and report data on students' understanding of CIs, and difficulties they have making appropriate interpretation of CIs. I explain the need for further evidence on which to base improved statistics education in psychology.

WHAT IS STATISTICAL REFORM?

There is clear evidence that researchers in psychology have many serious misconceptions about null hypothesis significance testing (NHST; Oakes, 1986). Haller and Krauss (2002) confirmed that the problems persist, and are also exhibited even by many teachers of statistics in psychology. Nickerson (2000) provided a review of NHST problems and of efforts by reformers to persuade psychologists to change their statistical practices.

Many reform advocates, including Jacob Cohen and Paul Meehl, have repeatedly warned against mindlessly replacing NHST with some other mechanical procedure. Their advice is good advice. The most needed improvement in any discipline is more thoughtful engagement with research data, and less reliance on automated decision making strategies. However, it is also unlikely that researchers' practice will change until a consensus is reached over a new approach to inference. The following are some common recommendations for statistical reform.

In psychology, especially in recent years, most reformers have advocated increased use of effect sizes and CIs, as either a supplement or a replacement for NHST (Cumming and Finch, 2001, 2005). Harlow (1997) identified CIs as the most commonly recommended alternative to NHST by contributors to *What If There Were No Significance Tests?* In medicine, CIs were also the most common recommendation. Statistical reporting in medical journals now largely reflects this consensus, with around 85% of 2003 articles in 10 leading medical journals reporting CIs (Coulson, Fidler and Cumming, in preparation). CIs have also received some—although admittedly less than in medicine or psychology—attention in the ecology reform literature.

Standardised effect sizes are also a commonly recommended supplement to NHST in psychology (Harlow, 1997). Being units-free, they have the virtue of facilitating meta-analysis. Meta-analysis itself is also a vital part of statistical reform (Cumming and Finch, 2001, 2005).

Another recommendation, common in ecology, is the use of information theoretic approaches—a move led by Ken Burnham, David Anderson and colleagues. Akaike Information Criteria (AIC), based on the work of Akaike (1992), has received particular attention. AIC is a likelihood-based model selection technique based on a trade-off between parsimony and fit. AIC is used to compare competing models, and to combine models to make multi-model inferences.

Bayesian methods have also received considerable attention in ecology, especially Bayesian model selection equivalents to the likelihood-based techniques just described. Bayesian methods have also had some strong, if proportionally fewer, advocates in medicine. In psychology, advocates of Bayesian methods have been around for decades, but they constitute a relatively small minority of NHST critics.

Whilst all or any of the practices listed above (effect estimation, CIs, standardised effect sizes, information theoretic and Bayesian methods) constitute statistical reform, so do more generic practices, such as: increased use of graphical representations; consideration of the clinical or biological or practical importance of results (as opposed to merely the statistical significance);

consideration of sample size issues; appreciation of meta-analytic issues and, more generally, the need for scientific results to be cumulative; and other thoughtful treatments of trends, patterns and effects that go beyond the mechanised dichotomous decision process of NHST.

My focus is on estimation and precision (i.e., effect sizes and CIs) as an alternative to NHST, and on why even minimal steps beyond NHST have been so difficult to achieve, and in many cases, are still a way off.

STATISTICAL REFORM EFFORTS IN THREE DISCIPLINES: MEDICINE, ECOLOGY AND PSYCHOLOGY

Medicine

In the 1950s medicine faced a flood of new 'wonder drugs' (Marks, 1997). Antibiotics and steroids were marketed for the first time and important decisions about their effectiveness had to be made quickly. Prior to this, the discipline had shown little interest in the statistical approaches of Fisher or Neyman and Pearson and rarely ran what could be considered randomised trials (Hogben, 1957, Part One). Therapeutic reformers—champions of the randomised clinical trial—were concerned that decisions made in the traditional way (that is, on the expert recommendation of individual physicians) were too time consuming and too open to biases and pressure from pharmaceutical companies. The then newly-arrived hypothesis testing techniques appeared to possess the qualities they were looking for: efficiency and objectivity.

Therapeutic reform was successful and NHST was rapidly institutionalised as a routine step in clinical trial procedure. By 1950 NHST was convincingly in use, if not widespread, in medicine—we found NHST was reported in 18% of *British Medical Journal* (*BMJ*), *New England Journal of Medicine* (*NEJM*) and *Lancet* articles in that year (Fidler, Cumming, Burgman and Thomason, 2004). A relatively systematic increase in reporting saw this rate reach 51% by 1970. A survey by Emerson and Coditz (1983) reported that, at the end of the 1970s, 44% of *NEJM* articles reported *p* values based on *t* tests and 27% computed *p* values from contingency tables.

But just over a decade after the institutionalisation of NHST in medicine, its role in clinical trials was under scrutiny. Researchers began to worry that the technique was being misused and over-relied on; that statisticians, rather than physicians, had authority over the conclusions drawn from experiments (Cutler *et al.*, 1966). Statistical reform had begun, and by the end of the 1980s strict editorial policies had profoundly changed the way results were reported, if not interpreted, in medicine.

Effect size and CI reporting became routine reporting practice in many journals. The most dramatic examples of policy change in medicine were Ken Rothman's editorial interventions, first at the *American Journal of Public Health* (*AJPH*) and later at *Epidemiology*. In 1986, as assistant editor of *AJPH*, Rothman wrote in revise-and-resubmit letters to authors: "All references to statistical hypothesis testing and statistical significance should be removed from the papers" (Rothman, cited by Fleiss, cited by Shrout 1997). In just a few years, sole reliance on *p* values dropped dramatically from 63% in 1982 (two years before Rothman's term as assistant editor began) to 6% in 1986 (two years after his term began; Fidler, Thomason, Cumming, Finch and Leeman, 2004). There was a concomitant increase in CI reporting: from 10% before Rothman, to 54% in 1986. At *Epidemiology* Rothman's policy was even stricter. He wrote: "when writing for *Epidemiology*, you can enhance your prospects if you omit tests of statistical significance… In *Epidemiology*, we do not publish them at all" (1998, p. 9). In 2000, over 90% of *Epidemiology* articles reported CIs and there were no *p* values, none at all (Fidler, Thomason, *et al*., 2004).

Many other leading medical journals underwent similar reforms in the mid 1980s (e.g., *BMJ, Medicine Journal of Australia*). The International Committee of Medical Journal Editors (ICMJE) lent their official support to the shift to CIs from NHST (ICMJE, 1989). New textbooks were quickly written to facilitate change. For example, Gardner and Altman (1989) identified a potential obstacle to the editorial reforms: "…the methods needed to calculate confidence intervals are not readily available in most statistical textbooks" (p. 4). They addressed this problem directly by writing such a textbook, *Statistics with Confidence*, which included dedicated software, *Confidence Interval Analysis*. Rothman (1988) offered *Modern Epidemiology*, an advanced statistical text that supported his CI approach to analysis. By contrast, in psychology

texts offering equivalent support to statistical reform have only been published in the past few years (e.g., Kline, 2004; Smithson, 2002; Thompson, in press; Zechmeister and Posavac, 2003).

Ecology

NHST reporting in ecology increased dramatically in the mid 1950s. In our survey of 1950 to 1970 issues of *Ecology* and *Journal of Ecology* NHST rose from just 6% (only 3 instances) in 1950, to 33% in 1955. From 1955, its reporting increased consistently to 60% in 1970 (Fidler, Cumming, Burgman and Thomason, 2004).

There has been little reason to think NHST lost favour amongst ecological researchers in the decades that followed the 1970s, at least until very recently. Despite growing criticisms of NHST in ecology, Anderson, Burnham and Thompson (2000) reported that "the estimated number of *p*-values appearing within articles of *Ecology* exceeded 8,000 in 1991 and has exceeded 3,000 in each year since 1984" (p.912). Similarly in the *Journal of Wildlife Management* there were over 3,000 *p* values a year from 1994 to 2000 (Anderson *et al*., 2000).

However, very recently it has been possible to detect some change in statistical reporting in ecology, or at least in the sub-discipline of conservation biology. In Fidler, Burgman, Cumming, Buttrose and Thomason (2005) we present results from a survey of 2001-2002 and 2005 issues of *Conservation Biology* and *Biological Conservation*. In 2001-2002, over 92% of empirical articles used NHST; in 2005, this figure had dropped to 78%. There were small but corresponding increases in the use of CIs, modelling techniques and Bayesian and Information Theoretic methods.

Whilst statistical reform efforts have a much shorter history in ecology than in other disciplines (the first wave of NHST criticisms did not enter this discipline until the 1980s), reform has come from more directions than in medicine or psychology. Early warnings about the consequences of ignoring statistical power began in the 1980s, but at that time few authors provided any other criticisms of NHST or advocated alternatives. More recently, advocates of Bayesian methods have been amongst the most outspoken critics; so too have proponents of likelihood and information theoretic methods. Unlike medicine and psychology, there has been relatively little attention paid to CIs in ecology—with the exception of DiStefano (2003) and a couple of others).

Whilst the criticisms of NHST are virtually identical to those made in medicine and psychology, in other respects the approach taken to statistical reform in ecology has been quite different. In ecology, Bayesian and information theoretic approaches will no doubt grow in popularity as their computational difficulties become less daunting, with faster computers and better-developed software.

Psychology

NHST was fully institutionalised in psychology journals, textbooks and training programs by the mid 1950s. Sterling (1959) surveyed articles in four leading psychology journals published in 1955 and 1956. An overwhelming 81.5% of these articles reported NHST. Hubbard and Ryan (2000) conducted a large survey of 12 American psychology journals. Like others (e.g., Hubbard, Parsa and Luthy, 1997) they found a systematic increase in the use of NHST over time, with NHST being reported in 94% of articles published between 1995 and 1998. Coulson, Fidler and Cumming's (in preparation) results confirm that this trend continues. In articles published in 2003, in 10 leading psychology journals, 98% of empirical articles reported NHST.

Such consistency in reporting may suggest this discipline has had no NHST critics. Yet, the statistical reform literature in this discipline is huge (my own Endnote library contains over 700 references for psychology). Furthermore, criticisms date back further than they do in either ecology or medicine. Psychology has also seen some committed editors attempt to motivate change through editorial policy (notably, Philip Kendall at *Journal of Consulting and Clinical Psychology*; Geoff Loftus at *Memory and Cognition*; and Bruce Thompson at *Educational and Psychological Measurement*). Yet these editorial initiatives (Kendall, 1997; Loftus, 1993; Thompson, 1994), like the published critiques before them, had little success in reforming practices. The Loftus initiative was assessed by Finch, Cumming, Williams *et al*. (2004) and the Kendall initiative by Fidler, Cumming, Thomason *et al*. (2005). The Thompson initiative has not

been similarly assessed, and may have had some ongoing effect at that journal. In general, however, any change achieved by editorial policy has been limited, short lived and largely failed to spread beyond individual journals. In 1996 the American Psychological Association (APA) Board of Scientific Affairs appointed a Task Force on Statistical Inference to investigate the ongoing controversy surrounding NHST and issue guidelines for statistical reporting (Wilkinson *et al*., 1999). To date these guidelines and the subsequent revision of the APA *Publication Manual* (APA, 2001)—which I subjected to a critical appraisal in Fidler (2002)—also appear to have had little impact on statistical reporting in the journals.

Amongst other complex sociological and philosophical explanations for psychology's resistance to reform is this fact: An estimation approach has been advocated (as a supplement or replacement to NHST) in many sciences, including psychology, for decades. Yet, few empirical questions have been asked about whether this new approach will be better understood, alleviate widespread misconceptions, or lead to more substantial interpretations of research findings. Do researchers and students understand CIs better than NHST? Can CIs avoid NHST misconceptions without introducing problems of their own? These empirical questions have to date received little research attention. Evidence on such issues is needed to guide design of the statistics education needed to support reformed statistical practice by researchers.

BUILDING AN EVIDENCE-BASE FOR REFORMED STATISTICAL EDUCATION

Empirical studies (with undergraduate ecology students from the University of Melbourne—all had at least one semester of introductory statistics) suggest that CIs help alleviate a particularly serious misconception associated with NHST, namely that statistical non-significance is equivalent to evidence of 'no effect.' When asked to interpret results presented in NHST format, 44% (24 of 55; 95% CI: 31 to 57%) of students misinterpreted statistically non-significant results—from a low powered study with a non-trivial effect size—as evidence that the null hypothesis is true. Less than half as many (18%, 10 of 55; 95% CI: 10 to 30%) made this mistake in the CI condition.

However, there is also less positive news! A second series of experiments (with 180 first or second year psychology and ecology students from La Trobe University and the University of Melbourne) revealed that CIs themselves are prone to a new set of unexpected and clear misconceptions. For example, many students failed to recognise CIs as inferential statistics, confusing them with the range of sample data (see Table 1). Other misconceptions associated with CIs went beyond fundamental confusion over their inferential nature (some are listed in Table 2).

Table 1: Percentage of students choosing each description of a CI from a prepared list

	% (*n* of 180)	95% CI
Plausible values for population mean*	22% (40)	17 to 29%
Plausible values for sample mean	38% (68)	31 to 45%
Range of individual scores	8% (14)	5 to 13%
Range of individual scores within one standard deviation	11% (20)	7 to 17%
Unsure	21% (38)	16 to 28%

*correct response; all others represent misconceptions—although 'unsure' may be accurate!

Table 2: Percentage of students agreeing with various statements about CIs

	% (*n* of 180)	95% CI
CI width decreases with sample size*	16% (28)	11 to 22%
CI width increases with sample size	20% (36)	15 to 26%
CI width unaffected by sample size	29% (52)	23 to 36%
Unsure of relationship between CI width and sample size	36% (64)	29 to 43%
90% CI wider than 95% CI (for same data)	73% (131)	66 to 79%

*correct response; all others represent misconceptions—although 'unsure' may be accurate!

CONCLUSION

It is too early to tell whether the misconceptions identified are due simply to CIs being unfamiliar. Perhaps with adequate training, better presentations and appropriate guidelines, such misconceptions would simply disappear? Whilst this question remains largely unanswered, this paper highlights the remarkable fact that, to date, statistical reform has been advocated—and in some disciplines even instituted—without an evidence base. One of the most compelling arguments against NHST is its tendency to be misinterpreted. If it is to be abandoned largely because of this, then surely the onus is on us to provide some evidence that whatever replaces it will be less frequently misunderstood—and further to ensure that it can be explained and taught to students.

REFERENCES

Akaike H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz and N. Johnson (Eds.), *Breakthroughs in Statistics*, (pp. 610-624). New York: Springer Verlag.

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

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

Coulson, M., Fidler, F., and Cumming, G. (in preparation). Understanding of confidence intervals by researchers in psychology, behavioural neuroscience, and medicine.

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

Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman M. A. (1966). The role of hypothesis testing in clinical trials. *Journal of Chronic Diseases*, 19, 857-882.

Emerson, J. D., and Colditz, G. A. (1983). Use of statistical analysis in the *New England Journal of Medicine*. *New England Journal of Medicine*, 312, 890-897.

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., and Thomason., N. (2005, accepted, pending revisions). Have criticisms of null hypothesis significance testing had an impact on statistical reporting practices in conservation biology? *Conservation Biology*.

Fidler, F., Cumming, G., Burgman, M. and Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio Economics*, 33, 615-630.

Fidler, F., Thomason, N., Cumming, G., Finch, S. and 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., and Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory and Cognition. Behavior Research Methods. *Instruments and Computers,* 36, 312-324.

Gardner, M. J. and Altman, D. G. (Eds.). (1989). *Statistics with Confidence.* London: *BMJ* Books.

Haller, H. and Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online,* 7(1).

Harlow, L. L. (1997). Significance testing in introduction and overview. In L. L. Harlow, S. A. Muliak and J. H. Steiger (Eds.). *What If There Were No Significance Tests?* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hogben, L. (1957) *Statistical Theory: The relationship of Probability, Credibility, and Error; An Examination of the Contemporary Crisis in Statistical Theory From a Behaviourist Viewpoint*. London: Allen and Unwin.

Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the *Journal of Applied Psychology,* 1917-1994. *Theory and Psychology*, 7, 545-554.

Hubbard, R. and 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.

Kline, R.B. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington DC: American Psychological Association.

Loftus, G. R. (1993). Editorial comment. *Memory and Cognition,* 21, 1–3.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods,* 5, 241-301.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* Chichester: Wiley.

Rothman, K. J. (1988). *Modern Epidemiology* (2nd edition). Philidelphia: Lippincott Williams and Wilkins.

Rothman, K. J. (1998). Writing for *Epidemiology. Epidemiology,* 9, 333–337.

Shrout, P. E. (1997). Should significance tests be banned? *Psychological Science,* 8, 1-2.

Smithson, M. (2002). *ConfidenceIintervals.* Thousand Oaks, CA: Sage.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association,* 54, 30-34.

Thompson, B. (in press). *Foundations of Behavioral Statistics: An Insight-Based Approach.* New York: Guilford.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement,* 54, 837-847.

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

Zechmeister, E. B. and Posavac, E. J. (2003). *Data Analysis and Interpretation in the Behavioral Sciences*. Belmont, CA: Wadsworth.