

RUNNING HEAD: Medicine, Psychology and Ecology

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

© Elsevier. This article may not exactly replicate the final version published in the journal. It is not the copy of record. For the *Journal of Socio-Economics* see:

http://www.elsevier.com/wps/find/journaldescription.cws_home/620175/description#description

Statistical Reform in Medicine, Psychology and Ecology

Fiona Fidler^{1,2,3}, Cumming, Geoff³, Burgman, Mark¹ and Thomason, Neil²

1 = Environmental Science, School of Botany, University of Melbourne, 3010 VIC

2 = History and Philosophy of Science, University of Melbourne, 3010 VIC

3 = School of Psychological Science, La Trobe University, 3086 VIC, Australia

Author contact:

Fiona Fidler

School of Botany, University of Melbourne, Victoria 3010, Australia

Phone: +61 3 8344 4405 Fax: + 61 3 9347 5460

Email: fidlerfm@unimelb.edu.au

Abstract

Over-reliance on Null Hypothesis Significance Testing (NHST) is a serious problem in a number of disciplines, including psychology and ecology. It has the potential to damage not only the progress of these sciences but also the objects of their study. In the mid 1980s, medicine underwent a (relatively) major statistical reform. Strict editorial policy saw the number of p values in journals drop dramatically, and the rate of confidence interval reporting rise concomitantly. In psychology, a parallel change is yet to be achieved, despite half a century of debate, several editorial inventions, and even an American Psychological Association Task Force on Statistical Inference. Ecology also lags substantially behind. The nature of the editorial policies and the degree of collaboration amongst editors are important factors in explaining the varying levels of reforms in these disciplines. But without efforts to also re-write textbooks, improve software and research understanding of alternative methods, it seems unlikely that editorial initiatives will achieve substantial statistical reform.

Over 30 years ago, Morrison and Henkel (1970) commented on the “parallel but quite independent scrutiny” (p. 182) Null Hypothesis Significance Testing (NHST) had undergone in different disciplines, even in the apparently closely related disciplines of psychology and sociology. This trend of independent scrutiny has largely continued, with various disciplines reinventing the controversy at different times over the last 60 years. This special issue of the *Journal of Socio-Economics* is a remarkable exception, in that its editor has actively sought an interdisciplinary context for the problem.

Medicine, psychology and ecology have had varying success in reducing misuse and misinterpretation of NHST. Medicine, for example, has made considerable progress in the way

results in academic journals are reported: Confidence Intervals (CIs) routinely replace or at least supplement p values. Psychology has produced many articles criticising NHST, and seen several editorial and institutional interventions, but journal surveys show only minimal improvement. Ecology is still in the nascent stages of reform: criticisms of NHST are growing, and alternatives have been suggested, but little editorial or institutional intervention has been attempted.

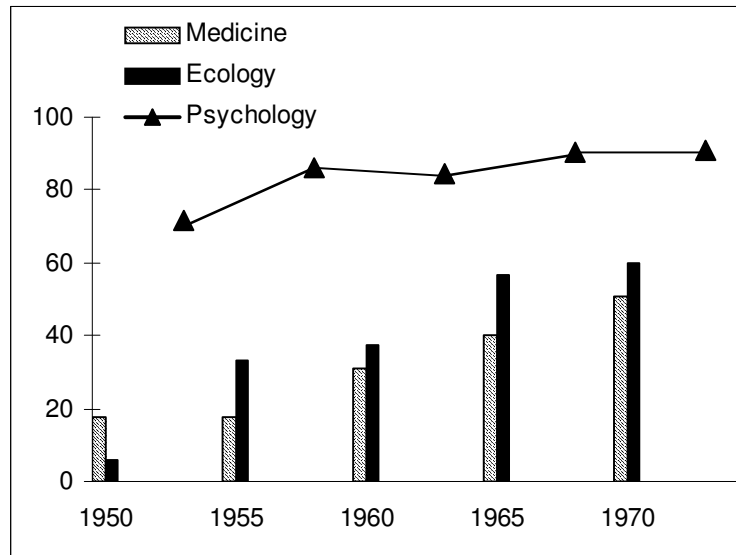


Figure 1. Percentage of articles using null hypothesis significance testing. Medical articles (total $n = 913$) in each of the years graphed from the *Lancet*, the *British Medical Journal* and the *New England Journal of Medicine*. Ecology articles (total $n = 524$) in each of the years graphed from *Ecology* and the *Journal of Ecology*. Psychology data from Hubbard and Ryan's (2000) survey of 12 APA journals. Shown here are percentages for 1950-1954 ($n = 431$), 1955-1959 ($n = 540$), 1960-1964 ($n = 609$), 1965-1969 ($n = 823$) and 1970-1974 ($n = 1014$).

Medicine

In medicine, the uptake of NSHT coincided with rise of the clinical trial (Figure 1 shows increasing use of NHST between 1950 and 1970). In the 1950s medicine faced a flood of new 'wonder drugs' (Marks, 1997). Antibiotics and steroids were marketed for the first time. Therapeutic reformers, champions of the clinical trial, were concerned that decisions made in the traditional way, on the expert recommendation of individual physicians, were too time consuming and too open to biases and pressure from drug companies. They believed hypothesis testing techniques possessed the qualities they were looking for: efficiency and objectivity. Their reform was successful and NHST was rapidly institutionalised as a routine step in clinical trial procedure. By the mid 1960s, however, the role of NHST in clinical trials was being questioned (e.g. Cutler et al, 1966). Researchers began to worry that the technique was being overused and that statisticians, rather than physicians, had authority over the conclusions drawn from experiments (Marks, 1997).

In the 1970s criticisms became increasingly common (e.g., Bandt & Boen, 1972; Schulman, 1976). Critics advocated reporting CIs in place of p values (e.g., Green, 1972; Wulff, 1973). To aid this transition, some provided guides to calculating CIs for relevant bio-medical effect sizes, such as odds ratios and relative risk values (e.g., Rothman, 1975, 1978a).

From virtually the beginning of debate in medicine, NHST reporting was a serious editorial concern. In 1977 the prestigious *New England Journal of Medicine (NEJM)* instigated a review of its own statistical reporting practices. The next year it published two editorials warning against the pitfalls of NHST and promoting the use of CIs (Rothman, 1978b, Rennie, 1978). The study group that conducted the *NEJM* review eventually produced more than 30 articles, and an edited book, *Medical Uses of Statistics* (Baliar & Mosteller, 1986).

In 1983 Ken Rothman became assistant editor of the *American Journal of Public Health (AJPH)*. Rothman had been on the editorial board of the *NEJM* in the late 1970s. Prior still he had published criticisms of NHST and 'how to' guides for calculating CIs (e.g. 1975). At *AJPH* he took the most radical stance yet been taken in statistical reform. In his revise and submit letters to would-be *AJPH* authors he wrote:

All references to statistical hypothesis testing and statistical significance should be removed from the papers. I ask that you delete p values as well as comments about statistical significance. If you do not agree with my standards (concerning the inappropriateness of significance tests) you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere. (Rothman, cited in Shrout, 1997, p.1)

Some years later, in 1990, Rothman became the founding editor of *Epidemiology*. Here his policy on NHST was similar:

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. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression (1998, p.9)

Rothman's contributions to reform are widely acknowledged (e.g. Altman, 2000). Fidler et al (2004) surveyed the effectiveness of Rothman's editorial policies, both at *AJPH* and *Epidemiology*. The increase in CI reporting at *AJPH* was dramatic: from 10% before Rothman to 54% during his editorship, and it remained high even after he left the journal. Sole reliance on p values dropped from 63% pre-Rothman to just 5%. Even more impressive is the impact of his policy at *Epidemiology*. For example, of 70 articles published in 2000, 94% reported CIs and none reported p values.

In 1982, the *British Medical Journal (BMJ)*, commissioned a series of articles by Douglas Altman and Shelia Gore on the use and misuse of statistics in medical practice. These were collected in book *Statistics in Practice* (1982). By 1986, *BMJ* had a policy recommending CIs rather than p values (Langman, 1986). And the percentage of articles reporting CIs consequently, increased from just 4% to 62% (Seldrup, 1997). In 1988, another *BMJ* editorial noted the spread of reform to other journals:

The *British Medical Journal* now expects scientific papers submitted to it to contain confidence intervals when appropriate. It also wants a reduced emphasis on the presentation of P values from hypothesis testing. *The Lancet*, the *Medical Journal of Australia*, the *American Journal of Public Health*, and the *British Heart Journal*, have implemented the same policy, and it has been endorsed by the International Committee of Medical Journal Editors. (Gardner and Altman, 1988, p.1210)

In fact, by 1988, over 300 medical and biomedical journals had notified the International Committee of Medical Journal Editors (ICMJE) of their willingness to comply with the guidelines for publication. On the matter of NHST, their guidelines instructed:

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).

A number of journal editors also involved themselves in the challenge of re-writing textbooks. Two texts are worthy of particular mention. Gardner and Altman's *Statistics with Confidence* was first published in 1989, by *BMJ* Books. (It is now in its second edition, Altman, Machin, Bryant & Gardner (Eds.), 2000). In their introduction, Gardner and Altman identified a serious obstacle to statistical reform, one psychology would have done well to note: "One of the blocks to implementing this policy [the editorial policy on statistical reporting at *BMJ*, cite earlier] has been that the methods needed to calculate confidence intervals are not readily available in most statistical textbooks" (1989, p.4). As the title suggests this text was CI based. Furthermore, it came complete with its own software: Confidence Interval Analysis (CIA). Ken Rothman's (1986) *Modern Epidemiology*, an advanced text, outlined problems with NHST and promoted CIs and p value functions as alternatives. It is now in its second edition (Rothman & Greenland, 1998).

In summary, statistical reporting in medicine was an editorial concern from the moment debate over NHST started. Editors of major journals strictly (some more strictly than others) enforced policy, important and relevant textbooks were written almost simultaneously, and substantial changes in reporting practice were evident by the end of the 1980s.

Psychology

Psychology embraced NHST almost immediately after WWII, and it quickly became popular, even more so than in medicine. In 1955, for example, just over 81% of articles in four leading empirical journals reported results of a significance test (Sterling, 1959); between 1955 and 1959, the percent of empirical articles with p values from 12 APA journals was 86% (Hubbard & Ryan, 2000; see Figure 1).

Why was this discipline in particular so seduced? In psychology, NHST was used in service of the emerging experimental psychology's ideals of objectivity and determinism (Gigerenzer, 1987). The new techniques helped psychology in its struggle towards scientific credibility. Of course, it has been long recognised that the dichotomous accept/reject decision outcome of NHST provides only the illusion of objectivity, particularly when type II error rates are often high and unknown, as is the case in psychology (e.g., Cohen, 1962; Rossi, 1990).

Although critics of NHST emerged a decade or so earlier than in medicine (e.g., Meehl, 1954; Rozeboom, 1960), editorial initiatives took about a decade longer. The first was in 1993 at *Memory & Cognition (M&C)*, then under the editorship of Geoff Loftus. Loftus encouraged the use of figures with error bars (either standard error or CI bars), and the omission of NHST (Loftus, 1993). His reform was only partially successful. During his term the number of articles with error bars rose by 34%—but even this large increase meant less than half of authors followed his recommendations (the proportion using figures with bars peaked at 41%) (Finch et al, 2004). And after Loftus left *M&C* the proportion of authors who followed his recommendations immediately began to fall: the reform was not sustained. Even when error bars were reported, they were almost always accompanied by NHST, which was used to interpret results.

There are now 23 psychology and education journals whose editorial policy encourages alternatives or at least warns about the pitfalls of NHST (Hill & Thompson, 2004). But journal surveys provide reasons not to be over-encouraged by this figure. For example, in 1997 Philip

Kendall, then editor of the *Journal of Consulting and Clinical Psychology*, encouraged authors to report clinical significance in addition to statistical significance. This resulted in only a trivial increase. In 1996, 36% (21 of 59) articles mentioned clinical significance; in 2000 and 2001, this figure was 40% (25 of 60) (Fidler et al, 2005).

Unfortunately, it is not only the policies of individual editors that have had little impact. Since the 1st edition in 1952, the *APA Publication Manual* has given advice on reporting tests of statistical significance. This advice has sometimes been problematic. For example, the 1st edition claimed: “Extensive tables of nonsignificant results are seldom required. For example, if only 2 of 20 correlations are significantly different from zero, the 2 significant correlations may be mentioned in the text, and the rest dismissed with a few words” (APA, 1952, p. 414). However, in 1994, the 4th edition took the positive step of recommending statistical power and effect sizes. Given that the *Manual* sets the editorial standards for over 1000 journals in psychology, education and related disciplines this should have been a great leap forward for reform. Yet a number of journal surveys, including Kirk (1996), Vacha-Haase, Nilsson, Reetz, Lance and Thompson (2000) and Finch, Cumming and Thomason (2001), concurred in finding little influence these recommendations.

In 1996, after ongoing pressure from NHST critics, the APA’s Board of Scientific Affairs established a Task Force on Statistical Inference (hereafter, Task Force) to investigate a proposal to ban NHST from APA journals. The Task Force stopped well short of banning NHST, but did produce clear and thoughtful guidelines on statistical reporting matters (Wilkinson et al, 1999). For example, they strongly emphasised the need for measures of effect size to be the primary outcome of a study, and the need to differentiate clinical or practical significance from statistical significance. They recommended increased use of figures with error bars, and discussed the advantages of CIs over p values.

In 2001, the *APA Manual* was again revised (a 5th edition), and its statistics guidelines updated, apparently to align with the Task Force report. Like the Task Force, the 5th edition recommended effect sizes (including some clinically relevant effect sizes), figures with bars, and CIs, which it called “the best reporting strategy” (p.22). However, neither the examples in the statistics section nor the template manuscript in the back of the *Manual* were revised to reflect the new recommendations. Researchers were given no guidance as to how CIs might be reported and interpreted, and no examples of CI reporting were given. Whilst it is still too early to test empirically, these deficiencies are likely to negate any positive effect the new recommendations may have had. For further discussion on problems with the *Manual’s* statistics guidelines see Fidler (2002) and Finch, Thomason & Cumming (2002).

In short, psychology has produced a mass of literature criticising NHST over the last five decades, including an extensive catalogue of the many associated cognitive fallacies and misconceptions, but there has been little improvement in reporting practices in journals. Even editorial policy and (admittedly half-hearted) interventions by the APA have failed to inspire any substantial change.

Ecology

NHST arrived in ecology a little later than in psychology or medicine. This can be seen in Figure 1, but more telling is that the first successful biometry text (Sokal & Rohlf) for this audience was published as late as 1969. The delay can perhaps be accounted for by obstacles randomisation. It is often impossible to construct anything like a randomised trial in ecology, particularly in conservation biology, where populations are often small and individuals may be cryptic and difficult to study. Despite such difficulties, NHST did eventually become common

place. Currently, it is extremely widely used in ecology and related areas: p values were reported in 92% (92 of 100) of 2000 and 2001 *Conservation Biology* and *Biological Conservation* articles (Fidler et al, 2004) and at similar levels in recent issues of *Ecology* and the *Journal of Wildlife Management* (Anderson et al, 2000).

Routine misinterpretation of NHST is serious in this discipline, and the neglect of statistical power is endemic. Fidler et al (2004b) found that, of the 92% of articles reporting p values in conservation biology journals, 81% (74 of 92) reported at least one statistically non-significant result. Only 3% — just 2 of the 74 articles with statistically non-significant results — reported statistical power! Yet almost half of them (47%, 35 of 74) interpreted the statistically non-significant result as evidence for ‘no effect’, ‘no impact’ or ‘no relationship’. Similar rates have been reported for other journals (e.g. Anderson et al, 2000; Peterman, 1990; Taylor and Gerrodette, 1993). The potential consequences of such errors in ecology and conservation biology are of serious concern. Low and unknown statistical power can lead to direct, unanticipated and unacceptable environmental damage (e.g. Parris & McCarthy, 2001; Taylor & Gerrodette, 1993). Such damage is often irreversible.

In ecology, criticisms of NHST emerged in the 1980s, decades later than in either medicine or psychology. Since then, most advocates of reform have focused on increased use of statistical power calculations (e.g., Fairweather, 1991; Green, 1989; Hayes & Steidl, 1997; Peterman, 1990; Mapstone, 1995; Taylor & Gerrodette, 1993; Toft & Shea, 1983). Confusion has been created, however, by recommendations to use post hoc or retrospective power analysis, based on the observed effect size. This practice has been severely and justifiably criticized (e.g., Hoenig & Heisey, 2001), but reform discussion in ecology has never moved convincingly to CIs. Unfortunately, the debate over power seems to have merely served to keep the NHST framework entrenched, without actually increasing the reporting rate of power calculations.

One promising step in ecology is a recent move towards the use of information theoretic approaches, led by Burnham, Anderson and others (e.g. Anderson et al, 2000; Burnham & Anderson, 2001; Spiegelhalter, Best, Carlin & van der Linde, 2002). Akaike Information Criteria (AIC), based on the work of H. Akaike (for review see Akaike, 1992), has received particular attention. AIC is a likelihood based model selection technique that is based on a trade-off between parsimony and fit. AIC may be used to compare competing models, and to combine, or average, models to make multi-model inferences. In addition to expository articles, there have been applications of these techniques in the literature (e.g. Frair et al, 2004; Gibson, Wilson, Cahill & Hill, 2004; Johnson, Seip & Boyce, 2004; Tyre et al, 2003). Whilst it is still too early to predict how widespread the uptake of AIC in ecology will be, the challenge of producing the textbook has already been met (Burnham & Anderson, 2002). Bayesian methods have also received considerable attention in ecology (e.g. Ellison, 1996; Harwood, 2000; Wade, 2000), though the extent of their uptake has not been formally surveyed.

Despite seemingly having greater reason to be concerned with over-reliance on and misuse of NHST, ecology has only relatively recently begun to engage with these problems. The current interest in information theoretic methods is promising, but serious work remains. To date there have been virtually no editorial or institutional interventions. As far as we are aware, only the *Journal of Wildlife Management* has published an editorial warning readers and would be authors of some pitfalls of NHST (The Wildlife Society, 1995). Unfortunately, it contained a number of conceptual errors and ambiguities (Otis, 1995).

On the persistent nature of NHST

The longevity of flawed practice has been attributed to the strongly intuitive nature of statistical fallacies associated with NHST. Tversky and Kahnman (e.g. 1971, 1982), for example, identified a number of these fallacies: the representativeness heuristic, neglect of prior probability and the misconception of randomness. Oakes (1986) famously used the inverse probability fallacy to explain why psychological researchers so infrequently reported statistical power. Schmidt and Hunter (1996) also use cognitive misconceptions to explain continued reliance on NHST.

That researchers do commit these fallacies is uncontroversial. Empirical evidence can be found in direct surveys of researchers' understanding (e.g. Tversky & Kahneman, 1971; Oakes, 1986) and in surveys of journals, where misinterpretations of p values are frequently identified (e.g. Vaache-Haase, 2000; Finch et al, 2001, Fidler et al, 2004a). Fallacies and misconceptions have no doubt played a major role in maintaining the statistical status quo. Because most researchers believe NHST tells them much more than it does (e.g. they believe the p value is a direct index of effect size, or that it is the probability that the null (or the alternative) is true (or false)), they are unwilling to give up the practice.

But if we accept the evidence that 20 years ago medicine achieved some substantial change, then 'intuitive fallacies and misconceptions' suddenly seem an insufficient explanation of the continuing use of NHST in other disciplines. There is no *a priori* reason why medical researchers would be less susceptible to the fallacies mentioned above. In fact, there is abundant evidence of misuse and misinterpretations of p values in medical journals prior to reform (e.g. Freiman, Chalmers, Smith & Kubeler, 1978; Godfrey, 1986). So why is it that other disciplines persist the technique? Were the critics in medicine somehow clearer than those in psychology and ecology? Did they make more compelling arguments? Were they published more often or in better journals than their psychology or ecology counterparts? The answer to all these questions certainly appears to be 'no': on these measures, the disciplines come up even. What, then, was present in medicine, facilitating reform, but absent from psychology and ecology? What can we learn from these cases about how to successfully instigate substantial disciplinary change?

Of course, medicine still faces challenges in fully institutionalising statistical reform (e.g. Fidler et al, 2005; Savitz, Tolo & Poole, 1994). Medical journals are far from a paradigm of best practice. However, they are comparatively free of sole reliance on p values compared to psychology or ecology journals. The difference between medicine and the other disciplines here is enough for the question to be genuinely interesting.

Disciplinary Differences

Textbooks. As Altman acknowledged, and we cited earlier, a serious obstacle to implementing the policy of CI reporting at *BMJ* was that explanations of and introductions to the new methods were not readily available in common textbooks. Gardner and Altman (1989) addressed this problem directly by writing such a textbook complete with dedicated software. Rothman (1988) did the same with *Modern Epidemiology*. In psychology the equivalent texts have only been published in the few years (e.g., Kline, 2004; Smithson, 2002; Zechmeister & Posavac, 2003). In ecology, Burnham and Anderson's text was first published in 1998, but not until the 2nd edition (2002) did it emphasise information theoretic approaches for multimodel inference.

Collaborations among editors. During the mid-1980s, editors of major medical journals (i.e., the International Committee of Medical Journal Editors, ICMJE) met annually to discuss,

amongst other things, statistical reporting in their journals. Through these meetings, the reforms spread from journal to journal rapidly, including the *Australian Journal of Medicine (AJM)* (Geoff Berry, then editor of *AJM*, personal correspondence, April 2003). Within a few years of each other all the major journals all had policies related to p value reporting, backed up by the ICMJE guidelines. This was perhaps a key ingredient in their success. In psychology, on the other hand, Geoff Loftus was attempting to enforce his policy on error bars at *M&C* in 1993 whereas the *APA Manual* didn't recommend CIs until 2001. As a lone editor he had only a limited and short-lived impact.

We might expect things to change for psychology when APA came on board. But even then, collaboration with editors was minimal. Editors of APA journals were not addressed by the APA's Task Force until the final guidelines for statistical reporting were published (Wilkinson et al, 1999) and the group was ready to disband. There was little reaction to the guidelines from editors, and no follow up meetings of the Task Force and editors (personal communication, Robert Rosenthal, co-chair of the Task Force, May 2003). In ecology there is currently no oversight committee, no vehicle to disseminate ideas, even if some individual editors were receptive and proactive. This may be partly explained by the disparate nature of the discipline: few researchers fully identify as ecologists, but rather think of themselves as conservation biologists, wildlife managers, risk assessors etc. Of course, this doesn't explain the lack of guidance within those specialities.

The nature of policy: requirements vs encouragements. Compare the editorial positions of Ken Rothman and Philip Kendall. In 2000, Rothman as editor of *Epidemiology* did not publish a single p value, and 94% of articles reported CIs (Fidler et al., 2004a). Kendall on the other hand had just 40% of authors following his encouragement to report clinical significance (Fidler et al, 2005). A major difference is the extent to which the policies were enforced. Kendall, unlike Rothman, did not reject papers that failed to follow the editorial advice: "a paper would not be rejected because of the absence of effect size data, but authors would be told of the need to report effect sizes." (Kendall, personal communication, April 9, 2001).

Admittedly, Rothman's policy was extreme even for medicine. Altman (2001) acknowledged this: "I am unaware of any other medical journal which has taken such a strong stance against p -values." (p.9) ShROUT (1997) called it a "virtual ban" (p.1). However, in general, policies in medical journals were much stricter than in psychology. For a start, they were often stated as *requirements*. For example, "The *British Medical Journal* now expects scientific papers submitted to it to contain confidence intervals..." (2001, p. 4).

For the most part, editors in psychology have, like Kendall, provided *encouragements* rather than *requirements*. (Bruce Thompson's policy at *Educational and Psychological Measurement* was an exception to this). The mere suggestion of requirements, bans or mandates related to statistical reporting have been met with overwhelmingly negative attitudes. The original proposal to ban p values was quickly dismissed by the TFSI. The *Manual* committee did not even broach the question. Some have described the idea of bans or requirements as impinging on researchers' intellectual freedom (see interviews reported in Fidler, 2002). Others are concerned, in the case of the *APA Manual* in particular, that not making requirements for statistical reporting sends a conflicting message: "To present an 'encouragement' [to report effect sizes and CIs] in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is the send the message, 'these myriad requirements count, this encouragement doesn't'" (Thompson, 1999, p. 162).

In medicine, there seems to have been much less debate over this process. There was some frustration expressed at the way Rothman instituted his changes—that the ban on p values was not discussed in an open forum (e.g., Fleiss, 1986). But Rothman explained that for him it was not an issue of intellectual freedom, just correcting mistakes as one would correct grammatical errors:

My revise and submit letters... were not a covert attempt to engineer a new policy, but simply my attempt to do my job as I understood it. Just as I corrected grammatical errors, I corrected what I saw as conceptual errors in describing data (K.J. Rothman, personal communication, July, 2002)

How to best interpret this difference between the disciplines is not straightforward. Is it a different vision of what statistical reform means? A mere fixing of errors, like correcting grammar (in medicine) versus a mandate that would seriously impinge on researchers' intellectual freedom (in psychology)? Or perhaps it should be interpreted as different understandings of the role of the journal editor?

Geoff Loftus' main interest in becoming editor of *M&C* was to bring about changes in statistical reporting (Geoff Loftus, personal communication, August 2001). So why not enforce the policy, instead publishing articles that failed to follow the recommendations? The answer is to be found in the remarkable resistance Loftus and his co-editors encountered. For example, Loftus had to personally calculate around 100 standard errors and CIs for authors who either didn't know how or refused to provide them (Finch et al, 2004). It is difficult to see how a blanket enforcement could have worked in this situation. Rothman, on the other hand, encountered little or no resistance in enforcing his policy, and certainly did not have to do these sorts of calculations himself (K.J. Rothman, personal communication, July, 2002). Enforcing editorial policy is no doubt a determining factor in the success of statistical reform. But how did medicine get to a position where it was possible to enforce policy, and how did it get there a decade before psychology's first attempt?

Consultation with statisticians and quantitative specialists. Virtually all medical schools employ at least one statistician—many have a biostatistical unit, or even a fully-fledged department. In the 1940s statisticians were mostly consulted after the fact with data to be 'fixed up'. In the 1950s and 1960s, they aligned themselves with advocates of clinical trials (Marks, 1997). Since the 1960s, medical researchers often consult with statisticians *a priori* about the design of studies. Similarly, most medical journals have statistical editors as well as substantive editors, and the scope of statistical reviewing is wide (George, 1985; Altman, 1991). This is rarely the case in psychology and ecology journals. Nor do these disciplines usually have a dedicated departmental statistician.

In 1982, the journal *Statistics in Medicine* was published for the first time. A roughly equivalent journal for psychology, *Psychological Methods*, started almost 15 years later, in 1996. *Psychological Methods* was an important step in psychology's reform. In some US Psychology departments, only content-based articles on a substantive (e.g. clinical, development, social) research topic were counted as professional publications in tenure applications. Articles on methodology and statistics in psychology experiments, published in methodology or statistics journals, were not rated. *Psychological Methods* provided an opportunity for psychologists to publish on statistics and methodology, without being penalised by the tenure process. In ecology the situation is worse because, unlike psychology even, most ecology programs do not include any formal statistical training.

Measurement and Estimation. Medicine's shift was from testing to estimation. Is estimation as the basis of psychology conceivable? One reason it is difficult to imagine is the lack of natural or even universally-agreed measurement units. In medicine, measurement scales are, for the most part, meaningful (e.g. number of deaths). At the very least, the scales are universal: everyone measures blood pressure in Hg/mm. In psychology, on the other hand, one study might measure anxiety using the Anxiety Scale Test; another might measure it by increases in heart rate or skin conductance. How can the two studies ever be compared? One often-suggested solution is to standardise the effect sizes, that is, report the effect in units of standard deviation. However, it is important to recognise that CIs for standardised measures are not straightforward. They typically require non-central distributions, something most psychologists have never encountered. Also, they involve highly intensive computing processes that have only recently been accessible to individual researchers (e.g. Cumming & Finch, 2001; Smithson, 2001).

The problem for psychology is therefore two-fold. First, because raw effect sizes in psychology often don't mean anything much, it is difficult to conceptualise a psychological science with estimation as its main focus. Secondly, once we move to the realm of standardised effect sizes, calculating CIs becomes a trickier enterprise.

These are not problems medicine encountered. The conceptual shift to estimation was straightforward and so were CI calculations for what became the effect sizes of choice, odds ratios and relative risk values. (Although units-free, odds ratios and relative risk values are not standardised and they still rely on the original scale of the study for full interpretation. In fact some reform advocates in medicine have argued that standardised effect sizes, such as Cohen's *d*, are invalid, e.g. Greenland, 1986, 1998.)

Comparatively, psychology's attempts to provide comprehensive CI methods for the statistics commonly used in its discipline were very late (e.g. Cumming & Finch, 2001; Fidler & Thompson, 2001; Smithson, 2001, 2002; Steiger & Foudali, 1997). This can perhaps be seen as part of a broader problem. Reform advocates in psychology have from time to time been admonished for relying on experimental scenarios that are over-simplified. Grayson, Pattison & Robins (1997) for example argued: "some recent attacks on significance testing in the psychological literature (e.g. Cohen, 1994; Hammond, 1996; Schmidt, 1996) have largely taken place in the context of simple models with few parameters." (p.69). In this sense, reformers themselves must also be held responsible for the lag in psychology. The good news for ecology is that this criticisms seems not to apply: information theoretic approaches have largely been advocated in the context of complex, real world problems.

As we indicated already, problems remain for medicine —statistical reform is far from complete. For example, Savitz, Tolo and Poole (1994) reported that in the *American Journal of Epidemiology* "the most common practice was to provide confidence intervals in results tables and to emphasize statistical significance tests in result text" (p.1047). Fidler et al (2004a) provided evidence that this continues: even when editorial pressure is maintained and CIs are reported (and NHST is not), authors still rely on NHST terms like 'significant difference' to interpret results.

Conclusion and Recommendations

The improvements in statistical reporting in medicine can be partially attributed to strictly enforced editorial policy, virtually simultaneous reforms in a number of leading journals and the timely re-writing textbooks to fit with policy recommendations. Similarly, psychology's lack of reform might explained by lone editors working in isolation, inconsistent advice from the APA

Manual and a lag in the re-writing of textbooks. In addition, the transformation to a science of estimation is perhaps itself a more difficult task for psychology, conceptually and computationally. Ecology's reform may be qualitatively different to either psychology or medicine.

One can only hope that when reform does occur in psychology and ecology, it will constitute more than superficial changes in journal reporting — that it will bring substantial changes in the way researchers think about measurement and uncertainty, rather than simply jumping editors' hurdles.

One reason editorial policy changes have been, and will continue to be, insufficient is because relevant knowledge is lacking. Little is known about how researchers think about CIs (for example) or what misconceptions might be associated with their use (Cumming & Finch, 2005). How are they best presented? Taught? Used to interpret research results? Similar information is needed about other alternatives to NHST, such as Bayesian methods and information theoretic methods. These are empirical questions and what has so far been conspicuously absent from reform debates, in any of these disciplines, is an evidence-based approach.

Acknowledgement

Thanks to Brendan Wintle for valuable comments, particularly on the ecology sections of this paper.

References

- Altman, D.G. (1991). Statistics in medical journals: Developments in the 1980s. *Statistics in Medicine*, 10, 1897-1913.
- Altman, D.G. (2000). Confidence intervals in practice. In D.G. Altman, D. Machin, T.N. Bryant & M.J. Gardner (Eds.) *Statistics with confidence* (2nd ed., pp. 6-14). London: BMJ Books.
- Altman, D.G., Machin, D., Bryant, T.N. & Gardner, M.J. (Eds.). (2000). *Statistics with Confidence* (2nd ed.). London: BMJ Books.
- American Psychological Association (1952). *Publication Manual of the American Psychological Association*. Washington, DC: Author.
- 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.
- Bailar, J.C., & Mosteller, F. (Eds.). (1986). *Medical Uses of Statistics*. Waltham, MA: NEJM Books.
- Bandt, C.L., & Boen, J.R. (1972). A prevalent misconception about sample size, statistical significance, and clinical importance. *Journal of Periodontology*, 43, 181-183.
- Burnham, K. & Anderson, D. (1998). *Model selection and inference*. New York: Springer-Verlag.
- Burnham, K. & Anderson, D. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies, *Wildlife Research*, 28, 111-119.
- Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inferences: A practical information theoretic approach* (2nd ed.). New York: Springer-Verlag.

RUNNING HEAD: Medicine, Psychology and Ecology

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.

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.

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.

Ellison, A. M. (1996). An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications*, 6, 1036-1046.

Fairweather, P.G. (1991). Statistical power and design requirements for environmental monitoring. *Australian Journal of Marine and Freshwater Research*, 42, 555-567.

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., Thomason, N., & Buttrose, R. (2004). Deficiencies in statistical reporting practices in conservation biology journals. *Manuscript in review*.

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.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, 61, 575-604.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, 61, 181-210.

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.

Fleiss, J.L. (1986). Significance tests do have a role in epidemiological research: Reaction to A.A. Walker. *American Journal of Public Health*, 76, 559-560.

Frair, J.L., Nielsen, S.E., Merrill, E.H., Lele, S.R., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B. & Beyer, H.L. (2004). Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology*, 41, 201-212.

Freiman, J.A., Chalmers, T.C., Smith, H., & Kuebler, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *New England Journal of Medicine*, 299, 690-694.

Gardner, M.J., & Altman, D.G. (1989). *Statistics with confidence*. London: BMJ Books.

- Gardner, M.J., & Altman, D.G. (1988). Estimating with confidence. *British Medical Journal (Clinical Research Edition)*, 296, 1210-1.
- George, S. L. (1985). Statistics in medical journals: A survey of current policies and proposals for editors. *Medical and Pediatric Oncology*, 13, 109-112.
- Gibson, L.A., Wilson, B.A., Cahill, D.M., & Hill, J. (2004). Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology*, 41, 213-223.
- Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. p.11-34. In Kruger et al. in full? (Eds.) *The Probabilistic Revolution (Vol.2)*. Cambridge, MA: MIT press.
- Godfrey, K. (1986). Comparing the means of several groups. In J.C. Bailar & F. Mosteller (Eds.) *Medical Uses of Statistics*. Waltham, MA: NEJM Books.
- Grayson, D., Pattison, P., & Robins, G. (1997). Evidence, inference, and the “Rejection” of the significance test. *Australian Journal of Psychology*, 49, 64-70.
- Green, M. (1972). Confidence limits. *The Lancet*, 2, 538.
- Greenland, S. (1998). Meta-analysis. In K.J. Rothman & S. Greenland (Eds.), *Modern Epidemiology* (2nd ed., pp. 643-673). Philadelphia: Lippincott-Raven.
- Greenland, S., Schlesselman, J., & Criqui, M. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, 123, 203-208.
- Harwood, J. (2000). Risk assessment and decision analysis in conservation. *Biological Conservation*, 95, 219-226.
- 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.
- Hayes, J.P., & Steidl, R.J. (1997). Statistical power and analysis and amphibian population trends. *Conservation Biology*, 11, 273-275.
- 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.
- Johnson, D.H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, 63, 763-772.
- Johnson, C.J., Seip, D.R. & Boyce, M.S. (2004). A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology*, 41, 238-251.
- Kendall, P.C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, 65, 3-5.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- Kline, R. B. (in press) *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.
- Langman, M.J.S. (1986). Towards estimation and confidence intervals. *British Medical Journal*, 292, 716.
- Loftus, G. (1993). Editorial comment. *Memory & Cognition*, 21, 1-3.

- Mapstone, B. D. (1995). Scalable decision rules for environmental impact studies: Effect size, Type I and Type II errors. *Ecological Applications*, 5, 401-410.
- Marks, H.M. (1997). *The progress of experiment: Science and therapeutic reform in the United States, 1900-1990*. Cambridge, UK: Cambridge University Press.
- 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. (1970). *The significance test controversy - a reader*. London: Butterworths.
- Oakes, M. W. (1986). *Statistical inference: a commentary for the social and behavioural sciences*. Chichester, U.K.: Wiley.
- Otis, D. (1995). Journal News. *Journal of Wildlife Management*, 59, 630.
- Parris, K. M., & M. A. McCarthy. (2001). Identifying effects of toe clipping on anuran return rates: the importance of statistical power. *Amphibia Reptilia*, 22, 275-289.
- Peterman, R.M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, 47, 2-15.
- Rennie, D. (1978). Vive la difference ($p < 0.05$). *New England Journal of Medicine*, 299, 828-829.
- Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58, 646-656.
- Rothman, K.J. (1975). Computation of exact confidence intervals for the odds ratio. *International Journal of Bio-Medical Computing*, 6, 33-39
- Rothman, K.J. (1978a). Estimation of the confidence limits for the cumulative probability of survival in life table analysis. *Journal of Chronic Disease*, 31, 557-560.
- Rothman, K.J. (1978b). A show of confidence. *New England Journal of Medicine*, 299, 1362-1363.
- Rothman, K.J. (1986). *Modern Epidemiology*. Boston, MA: Little and Brown.
- Rothman, K.J. (1998). Writing for Epidemiology. *Epidemiology*, 9, 333-337.
- Rothman, K.J. & Greenland, S. (Eds.). (1998). *Modern Epidemiology* (2nd ed.). Philadelphia, PA : Lippincott-Raven.
- Rozeboom. W.W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Savitz, D.A., Tolo, K., & Poole, C. (1994). Statistical significance testing in the *American Journal of Epidemiology*, 139, 1047-1052.
- Schulman, J.L., Kupst, M.J., & Suran, B.G. (1976). The worship of "p": significant yet meaningless research results. *Bulletin of the Menninger Clinic*, 40, 134-43.
- 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.
- Seldrup, J. (1997). Whatever happened to the t-test?. *Drug Information Journal*, 31, 745-50.
- Shrout, P. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, 8, 1-2.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Methods*, 61, 605-632.
- Smithson, M. (2002). *Confidence Intervals*. Thousand Oaks, CA: Sage.
- Sokal, R.R., & Rohlf, F.J. (1969). *Biometry*. New York: Freeman.

Steiger, J. H., & Foudali, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 222-257). Hillsdale, NJ: Erlbaum.

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.

Taylor, B. L., & Gerrodette, T. (1993). The use of statistical power in conservation biology: the Vaquita and northern spotted owl. *Conservation Biology*, *7*, 489-500.

Toft, C. A., & Shea, P.J. (1983). Detecting community-wide patterns: estimating power strengthens statistical inference. *American Naturalist*, *122*, 18-25

Tversky, A., & D. Kahneman. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105-110.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982) *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology*, *10*, 413-425.

Wade, P. R. (2000). Bayesian methods in conservation biology. *Conservation Biology*, *14*, 1308-1316.

The Wildlife Society. (1995). Journal news. *Journal of Wildlife Management*, *59*, 630.

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

Zechmeister, E. & Posavac, E. (2003). *Data Analysis and Interpretation in Behavioral Sciences*. Belmont, CA: Thomson.

Wulff, H.R. (1973). Confidence limits in evaluating controlled therapeutic trials. *The Lancet*, *2*, 969-970.