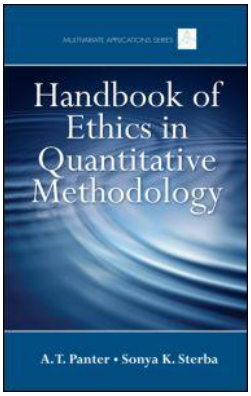

Handbook of Ethics in Quantitative Methodology

A.T. Panter, Sonya K. Sterba

Ethics and Statistical Reform: Lessons From Medicine

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203840023.ch17
Fiona Fidler
Published online on: 20 Jan 2011

17

Ethics and Statistical Reform: Lessons From Medicine

Fiona Fidler
La Trobe University

In psychology, *null hypothesis significance testing* (NHST; Cumming & Fidler, Chapter 11, this volume) and *meta-analysis* (MA) (Cooper & Dent, Chapter 16, this volume) have occupied advocates of *statistical reform* for decades. Hundreds of psychology journal articles criticize the former and encourage more widespread use of the latter. In medicine, NHST has similarly been admonished and MA promoted. Misuse and misinterpretation of NHST have been widespread in both disciplines. The alternative statistical practices advocated by reformers have been the same in both disciplines, too—estimation (*effect sizes* and *confidence intervals* [CIs]) and, increasingly, MA.

Despite these similarities between the disciplines, changes to statistical practice have been much slower in psychology than in medicine. For example, in 2006, 97% of articles in 10 leading psychology journals still reported NHST as the primary outcome (Cumming et al., 2007). In medicine, by contrast, CIs replaced NHST as the dominant analysis in individual studies in the mid-1980s (Fidler, Thomason, Cumming, Finch, & Leeman, 2004), and they remain a routine feature, being reported in approximately 85% of empirical articles (Cumming, Williams, & Fidler, 2004). In medicine, editorial policy in leading journals (e.g., *BMJ, The Lancet*) now requires that all new trials be placed in the context of previous research and integrated using MA (Young & Horton, 2005). Systematically placing new empirical results in the context of existing quantitative data is far from routine practice in psychology, although MA is certainly increasing.

The dramatic shift from rare to routine reporting of CIs in medical journals in the 1980s was supported by strict individual editorial policies and the institutional support of the International Committee of Medical Journal Editors (Fidler et al., 2004). I have argued elsewhere (Fidler, 2005) that the relative success of medicine's statistical reform occurred partly because medicine framed these statistical issues in ethical terms. Psychologists

and other behavioral scientists, on the other hand, presented mainly only technical and philosophical reasons for the advocated change. *Statistical power*, effect sizes, CIs, and other reform statistics were no longer merely technical issues to be worked out on a calculator or in an analysis software package or relegated to the consultant brought in after data collection, nor were they merely philosophical problems about the nature of evidence or the interpretation of probability. Rather, statistical reform was a practical and ethical concern, with obvious and tangible consequences, for every researcher, statistician or not. In psychology, this context has been largely lacking, the current edited volume being an obvious exception.

In this chapter, I first explicate how an ethical imperative was explicitly used in medicine to discourage NHST and to encourage MA. Next, I discuss two case examples from medicine that have been used to illustrate to practitioners why misuse of these techniques has clear ethical implications. I then provide two parallel examples from psychology that have similar—although comparatively underappreciated—ethical implications. Finally, I discuss reasons why an ethical imperative has been, to date, used in medicine but not psychology, why this is a problem, and how it can be remedied. In so doing, this chapter addresses several questions: Why were the ethical implications of statistical practice so salient to medical reformers but not psychological ones? What gains, in terms of statistical reform, did an ethical imperative afford medicine? What lessons can psychology learn from medicine's reform efforts, as well as from its mistakes?

In Medicine, Statistical Inference Is an Ethical Concern

One of the main criticisms of typical NHST practice in both medicine and psychology has, over the decades, been the neglect of statistical power. Calls for increased attention to statistical power have been the focus of hundreds of articles in both disciplines. The type of arguments used to promote statistical power, however, provides one of the clearest demonstrations of the differences between the disciplines.

In medicine, neglect of statistical power was identified as an ethical problem from early in the reform process. This is evident in the medical literature of the 1970s, as the following quotations demonstrate:

> One of the most serious ethical problems in clinical research is that of placing subjects at risk of injury, discomfort, or inconvenience in experiments where there are too few subjects for valid results. (May, 1975, p. 23)

> Not every clinician—or even his ethical committee—is acutely attuned to the details of statistical Type II errors. (Newell, 1978, p. 534)

In psychology's reform literature, by contrast, an ethical argument for statistical power has rarely been made explicit. Instead, we have seen analysis of the reporting rates of power (e.g., Fidler et al., 2005; Finch, Cumming, & Thomason, 2001); calculations of the average power of research (e.g., Cohen, 1962; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989); studies of misconceptions about power and sample size (starting with the law of small numbers; Haller & Krauss, 2002; Oakes, 1986; Tversky & Kahneman, 1971); technical explanations of statistical power; and finally philosophical explanations for the neglect of power (e.g., various authors refer to NHST as an incoherent amalgamation of Fisher and Neyman–Pearson, most notably Gigerenzer, 1993). All of these discussions are important in their own right, but none necessarily deals with the ethics of our current practice of neglecting Type II errors.

The ethical framing of this issue within the medical discipline was not an afterthought, nor was it a last-ditch rhetorical effort—rather, it was the primary impetus for statistical reform 3 decades ago. Altman (1982a) explains why:

> A study with an overly large sample may be deemed unethical through the unnecessary involvement of extra subjects and correspondingly increased costs. Such studies are probably rare. On the other hand, a study with a sample size that is too small will be unable to detect clinically important effects. Such a study may thus be scientifically useless, and hence unethical in its use of subjects and other resources. (Altman, 1982a, p. 6)

In the following quotation, Altman (1982b) spells out the consequences of neglecting statistical power:

> (1) The misuse of patients by exposing them to unjustified risk and inconvenience; (2) the misuse of resources, including the researchers' time, which could be better employed on more valuable activities; and (3) the consequences of publishing misleading results, which may include the carrying out of unnecessary further work. (Altman, 1982b, p. 1)

In medicine, particular cases in which flawed statistical practice continued, such as the ongoing neglect of statistical power or lack of attention to effect sizes, became scandals. The attention given to the neglect of statistical power was not seen as statistical nitpicking, but rather as justified criticism of professional misconduct. Here, Altman (1994) encourages researchers to be outraged when they come across misuse of statistics:

> What should we think about a doctor who uses the wrong treatment, either willfully or through ignorance, or who uses the right treatment wrongly (such as by giving the wrong dose of a drug)? Most people would agree that such behaviour was unprofessional, arguably unethical, and certainly unacceptable. What, then, should we think about researchers who use the wrong techniques (either willfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled.… This is surely a scandal. (Altman, 1994, p. 283)

In Medicine, Meta-Analysis Is an Ethical Concern

The ethical imperative for MA was also explicit in medicine, and its neglect also identified as wasting valuable research time and resources (e.g., the title of the article this quotation is from is "The Scandalous Failure of Science to Cumulate Evidence Scientifically"):

> New research should not be designed or implemented without first assessing systematically what is known from existing research.… The failure to conduct that assessment represents a lack of scientific self-discipline that results in an inexcusable waste of public resources. In applied fields like health care, failure to prepare scientifically defensible reviews of relevant animal and human data results not only in wasted resources but also in unnecessary suffering and premature death. (Chalmers, 2005, p. 229)

In 2005, *The Lancet* made MAs a requirement—new trials must be put in the context of previous research. This innovation was also justified on ethical grounds, with this explicit statement that continuing trials without conducting an MA is both unscientific and unethical:

> The relation between existing and new evidence should be illustrated to an existing systematic review or meta-analysis. When a systematic review or meta-analysis does not exist, authors are encouraged to do their own.… Those who say systematic reviews and meta-analysis are not "proper research" are wrong; it is clinical trials done in the absence of such reviews and meta-analysis that are improper, scientifically and ethically. (Young & Horton, 2005, p. 107)

Perhaps the best example of institutional acceptance of MA in medicine is the Cochrane Collaboration (http://www.cochrane.org), which is dedicated solely to conducting MAs of clinical trials to improve health care.

The Cochrane Collaboration was established in 1993 and has since published thousands of MAs and has clinical trial centers around the world. The Collaboration itself grew out of an ethical concern. Archie Cochrane's (1972) *Effectiveness and Efficiency* laid out the basic principle: Because health care resources would always be limited, the only ethical system was one that practiced only those treatments for which evidence had accrued from systematic, rigorous evaluation. Five years later, Cochrane (1979) laid the final challenge with this now famous quotation: "It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials" (p. 1). These words became a rallying point at the foundation of the Cochrane Collaboration (Chalmers, 2006).

A promising social science parallel, the Campbell Collaboration, began in 1999 (http://www.campbellcollaboration.org). It grew out of the Cochrane Collaboration and shares the goal of increased efficiency through evidence-based decision making. The Campbell Collaboration specializes in meta-analytically reviewing evidence related to education, crime, justice, and social welfare. Unfortunately, this still leaves a lot of clinical and experimental psychology territory uncovered. Another recent MA development in psychology is the Meta-Analytic Reporting Standards (MARS) section in the sixth edition of the American Psychological Association (APA) *Publication Manual* (2010) (see Cooper & Dent, Chapter 16, this volume). However, despite its excellent content, MARS is a mere appendix to the *Manual* and may be easily missed by the casual reader.

In the next section, I outline examples of studies where misinterpretations of NHST have led medical research astray and resulted in both a waste of resources and unnecessary suffering. These case studies also illustrate the importance of cumulative MA in sorting out the confusion left by dichotomous accept–reject decisions made in single experiments. In a subsequent section, I show that parallel examples in psychology also exist but have been less publicized and thus far had less impact on the statistical reform of the discipline.

Two Medical Case Examples

Medical Case Example 1: Myocardial Infarction and Streptokinase

Streptokinase is an enzyme that dissolves vascular thrombi, or blood clots caused by atherosclerosis. In the 1950s, medical researchers began to wonder whether it might benefit acute myocardial patients because most cardiac arrests are caused by atherosclerosis—a gradual buildup of a fat-containing substance in plaques that then rupture and form blood

clots on artery walls. Between 1959 and 1988, 33 randomized clinical trials tested the effectiveness of intravenous streptokinase for treating acute myocardial infarction. The majority of these trials (26 of 33) showed no statistically significant improvement at $p < .05$. However, the remaining trials did show a statistically significant improvement, and often a dramatic one. Those improvements were enough to motivate testing to continue, in pursuit of a definitive answer.

If one looks at the results of these trials as CIs, rather than as simply statistically significant or not, it is immediately obvious that those nonsignificant trials have extremely wide CIs. (An excellent graphic can be found in Lau et al., 1992, reprinted with copyright permission in Hunt, 1997.) The CIs of the statistically nonsignificant trials do indeed capture the odds ratio of 1—but they also capture almost every other value on the scale! Wide intervals are an immediate sign that the nonsignificant trials had low statistical power. The seemingly inconsistent results were a simple product of relative power of the trials. The high-powered trials produced statistically significant results; the low-powered trials (in this case, those with small sample sizes) did not.

In 1992, Lau et al. demonstrated that cumulative odds ratio (i.e., the odds ratio produced by an MA after the first two trials, another after the first three trials, and so on) was consistently greater than 1 by the time the fourth clinical trial was added and that the CI around this odds ratio did not capture 1 from the time the seventh clinical trial was added. Recall that there were 33 clinical trials—this result means that there were at least 26 more than there should have been! Overreliance on dichotomous accept–reject decisions from NHST, as well as neglect of statistical power, resulted in the unnecessary testing of 30,000 additional patients over an extra 15 years, half of whom were in the placebo group and therefore denied a treatment already proven to be effective. Presenting the results of individual trials with CIs—and placing those individual trials in the context of previous research by using cumulative MA—would have clearly shown that evidence in favor of the drug was indisputable and that additional subjects and years of further research were redundant.

In a separate publication of the same year, the same team of researchers who demonstrated and emphasized the unethical failings of the above study (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992) presented a comparison of textbook advice on the treatment of people with myocardial infarction and the results of several cumulative MAs. In each case, they showed that advice on lifesaving treatments had been delayed for more than a decade, and, in some, that harmful interventions were promoted long after evidence of their damage had accumulated.

The reports of these researchers on treatment for myocardial infarction have been identified in the medical literature as a "great impetus"

in the widespread recognition of the practical and ethical importance of unbiased, quality scientific reviews: [They] "made it abundantly clear that the failure of researchers to prepare reviews of therapeutic research systematically could have very real human costs" (Chalmers, Hedges, & Cooper, 2002, p. 21). In the same year, the Cochrane Collaboration was born.[1]

Medical Case Example 2: Antiarrhythmic Drugs

Antiarrhythmic drugs suppress the fast rhythms of the heart and were often prescribed after a cardiac arrest to prolong life. Many clinical trials assessed their safety, and although results were somewhat mixed, the accepted conclusion on an individual study basis was that there was, at worst, no difference in the mortality rate when the drugs were prescribed. This conclusion of "no difference" unfortunately turned out to be an overly optimistic interpretation of the research when an MA was performed. In 1993, an MA was carried out, examining 51 trials of Class I prophylactic antiarrhythmic agents conducted on 23,229 patients (Teo, Yusif, & Furberg, 1993).[2] The results clearly showed a substantially increased mortality rate as a result of the drug in question. Within the drug group there were 660 deaths (5.63% of patients) as opposed to 571 deaths in placebo groups (4.96% of patients).

The ethical implications of this case were swiftly picked up by commentators in the medical literature. Most famously, Moore (1995) commented that the number of deaths from these antiarrhythmic drugs at the peak of their use (in the late 1980s) was comparable with the number of Americans who died in the Vietnam War. Chalmers also discussed the ethics of this case and again argued explicitly that cumulative MA could have saved these lives (e.g., Chalmers, 2005). These comments lent timely support to the argument for the practical importance of cumulative MA, established by the myocardial case above.

---

[1] The other important development in the establishment of the Cochrane Collaboration was a large-scale synthesis of studies relating to pregnancy and childbirth. Chalmers, who was the lead author of the report and later became the founding leader of the Cochrane Collaboration, explained: "International collaboration during this time [the late 1980s] led to the preparation of hundreds of systematic reviews of controlled trials relevant to the care of women during pregnancy and childbirth. These were published in a 1,500-page, two-volume book, *Effective Care in Pregnancy and Childbirth* (Chalmers, Enkin, & Keirse, 1989), deemed an important landmark in the history of controlled trials and research synthesis (Cochrane, 1989; Mosteller, 1993)." (Chalmers, Hedges, & Cooper, 2002).

[2] The meta-analysis also looked at other types of agents, including beta blockers or Class II agents (55 trials), amiodarone or Class III agents (8 trials), or calcium channel blockers or Class IV agents (24 trials), but it is the results of Class I agents that are of particular interest here.

Comparing Medical Versus Psychology Case Examples

One could be reasonably skeptical about whether such strong linkages between statistical practice and ethics could be forged in psychology, as were achieved in the aforementioned two medical case examples. There may be complicated political and financial pressures associated with pharmaceutical trials that do not apply in psychological research. One could argue, for example, that medical statistics have to be stricter than psychology statistics to prevent unethical behavior by Big Pharma. However, not all cases of cumulative MA from medicine are drug trials. A more recent cumulative MA showed that advice on infant sleeping position, namely, to put newborns to sleep on their backs, was delayed for decades because of a failure to properly assess cumulative results from individual trials.

Advice to put infants to sleep on their fronts for nearly half a century was contrary to evidence available from 1970 that this was likely to be harmful. Systematic review of preventable risk factors for sudden infant death syndrome (SIDS) from 1970 would have led to earlier recognition of the risks of sleeping on the front and might have prevented more than 10,000 infant deaths in the United Kingdom and at least 50,000 in Europe, the United States, and Australasia (Gilbert, Salanti, Harden, & See, 2005).

Gilbert et al. (2005) make practical and ethical implications clear by discussing the consequences of the statistical error in number of thousands of infant deaths—and yet there is no pharmaceutical conspiracy to blame in this case.

However, an extensive search of the reform literature in psychology for equivalently explicit ethical discussions of statistical power and/or misuse of NHST, as well as for explicit framing of consequences of errors in practical terms of damage done (e.g., lives lost, resources wasted, or equivalent harm), was not fruitful. Although some articles infer ethical consequences of poor statistical practice (e.g., Meehl's 1978 article is perhaps the best example), none primarily emphasizes the ethical dimension of these problems as do the case examples in the medical literature. When we hear that psychologists misuse statistical techniques or misinterpret their results, are we "appalled," as Altman urged medical researchers to be? Does a scandal erupt, as it does in the medical literature? Rarely.

And yet, it is possible to find similar (albeit, less publicized) examples of irresponsible statistical analysis from psychology with important ethical implications. Indeed, I present two examples in the following section. These cases demonstrate that it is not simply the medical field's political and financial pressure that leads to ethically concerning statistical practice. The consequences of the poor practice remain an ethical concern, regardless of the discipline. I propose that the fact these psychological cases are comparatively unknown and have not been identified in the

psychology literature, or even in the reform literature, is in itself a scandal of alarming proportions. Additionally, their lack of publicity represents a missed opportunity for psychologists to add an ethical imperative to statistical reform.

Two Psychology Case Examples

Psychology Case Example 1: Employment Testing and the Theory of Situational Specificity

Schmidt claimed, as APA Division 5 President in 1996, that "reliance on statistical significance testing … has systematically retarded the growth of cumulative knowledge in psychology" (p. 115). Schmidt's dramatic quotation is now famous. However, I suspect the evidence for it is less well known, much less regarded as a scandal. Hunter and Schmidt (2004) themselves obviously consider the theory of situational specificity (TSS) scandalous (as do I), but it has rarely been advertised as such, even in the statistical reform literature. The evidence is a series of MAs that Schmidt, Hunter, and others conducted throughout the 1970s and early 1980s. These MAs exposed a great flaw in the then orthodox doctrine of organizational psychology.

The TSS pertains to employment tests—professionally developed cognitive ability and aptitude tests that are designed to predict job performance. The theory holds that the correlation between test score and job performance does *not* have general validity; that is, "a test valid for a job in one organization or setting may be invalid for the same job in another organization or setting" (Schmidt & Hunter, 1981, p. 1132). The validity of the tests, it seemed, depended on more than just the listed tasks for a given position description. The theory proposed that the validity of any one test depended on the cognitive information-processing and problem-solving demands of the job and perhaps even the social and political demands of the workplace. In other words, TSS proposed that there is a distinct context for each job, and that a general employment test may not predict specific job context performance.

How did the TSS come about? The belief in "situational specificity" grew out of the considerable variability observed from study to study, even when the jobs and/or tests were similar. Some studies found statistically significant correlations, whereas others found none. "Situational specificity" explained the inconsistency in the statistical significance of empirical results by generating potential moderating variables. Another obvious factor that could also explain why one study found a statistically significant result and another study did not was the varied and usually

low statistical power of the studies. This alternative explanation, however, went unnoticed for several decades.

The TSS grew structurally complex, with addition of many potential moderating variables, including organization size, gender, race, job level, and geographic location. In fact, the search for such moderating variables became the main business of industrial or organizational psychology for decades. Researchers sought to shed further light on the "specific" nuances of the theory, despite the fact that the variability that they were working to explain was illusory. Not until 1981, when Hunter, Schmidt, and their colleagues carried out an MA using the results of 406 previous studies, did it finally become clear that the difference in allegedly inconsistent results could be exclusively accounted for by the low statistical power of the studies.

> If the true validity for a given test is constant at .45 in a series of jobs … and if sample size is 68 (the median over 406 published validity studies…) then the test will be reported to be valid 54% of the time and invalid 46% of the time (two tailed test, p = .05). This is the kind of variability that was the basis for theory of situation-specific validity. (Schmidt & Hunter, 1981, p. 1132)

As Schmidt and Hunter finally revealed, the reporting of individual results as "significant" or "nonsignificant" had created the illusion of inconsistency, even though almost all the obtained effect sizes were in the same direction.

How long did organizational psychology pursue this misdirected theory and its associated research program? In 1981, toward the end of their MA series, Hunter and Schmidt wrote: "the real meaning of 70 years of cumulative research on employment testing was not apparent [until now]" (p. 1134). Of the use of NHST in this program, they wrote: "The use of significance tests within individual studies only clouded discussion because narrative reviewers falsely believed that significance tests could be relied on to give correct decisions about single studies" (p. 1134).

The case of the TSS provides evidence that NHST, as typically used—with little regard for statistical power and overreliance on dichotomous decisions—can seriously damage scientific progress. In this case, an important research program was led astray by a search for moderating variables to explain illusory differences. Years of empirical data were seen to support a theory, for which there was, in fact, no empirical evidence. A program of this scale going astray represents an enormous waste of public funds and scientific resources, including person hours, dollars, careers, and other research not conducted at the expense of this program and/or because the real findings were obscured. These losses are themselves serious ethical concerns. However, Schmidt and Hunter (1981) also hint at another, perhaps more disturbing, level of damage:

> Tests have been used in making employment decisions in the United States for over 50 years.… In the middle and late 1960s certain theories about aptitude and ability tests formed the basis for most discussion of employee selection issues, and in part, the basis for practice in personnel psychology.… We now have.… evidence.… that the earlier theories were false. (pp. 1128–1129)

In other words, the false TSS findings influenced the success of companies that relied on it (including missing out on potentially valuable employees who were rejected on the basis of test results, and vice versa), not to mention the careers of uncounted jobseekers. Despite the disturbing ethical implications of Schmidt and Hunter's findings, their debunking of the TSS failed to motivate widespread statistical reform regarding MA in psychology. Unlike the parallel cases in medicine—whose ethical implications helped launch the Cochrane Collaboration—this less publicized scandal about employment test validity had strikingly little impact on statistical practice in psychology.

Psychology Case Example 2: Learned Helplessness and Depression

The second psychology case example concerns *learned helplessness*, a concept pioneered by Seligman. The phenomenon was first isolated in dogs (Seligman, Maier, & Geer, 1968), much in the tradition of Pavlov. Caged dogs were given random electric shocks from which they could not escape. Later they were placed in different cages with separate compartments that they could use to escape from the shocks. They were again administered shocks. Surprisingly, around two thirds of the 150 dogs did not try to escape. They remained in the shock compartment and did not attempt to move. Seligman concluded that the dogs had learned that they were helpless. Immediately, Seligman and his colleagues began to wonder what links learned helplessness (or pessimistic explanatory style) might have with depression and illness.

Throughout the 1970s and early 1980s, strong links between pessimistic explanatory style–learned helplessness and depression and illness soon were made. For example, the effects of helplessness on growth of cancerous tumors and death rates were first observed in rats, and later experiments demonstrated the links in human subjects. Seligman and his colleagues published at least 25 articles on the topic between 1969 and 1977, as well as a book *Helplessness: On Depression, Development and Death* (Seligman, 1975). However, other researchers had trouble replicating the experimental results linking explanatory style to depression—or rather, they had trouble replicating the statistical significance of the results (see the 1978 special issue of the *Journal of Abnormal Psychology*). Eventually, several MAs showed that the inconsistencies in the literature were an artefact of NHST.

Specifically, an MA by Sweeney, Anderson, and Bailey (1986) combined 104 studies, excluding those from Seligman's lab, and for the first time found results consistent with Seligman's. Second, a series of statistical power analyses by Robins (1988) pointed out that only 8 of 87 previous individual studies on depression and explanatory style (or "attributions" as Robins calls them) had an a priori power of .80 or better for detecting the small population effect. Robins explained that the situation was so poor that "even adopting the assumption of a larger true effect, which I term *medium* (e.g., $r = .30$), only 35 of the 87 analyses had the desired chance of finding such an effect" (p. 885).

In sum, the misinterpretation of statistically nonsignificant results produced by underpowered learned helplessness studies caused several decades of ongoing debate and confusion where there should have been none. The academic damage in this instance could have been that a valid theory was lost for all time. Still, for at least a decade, important theoretical developments and clinical interventions based on relationships between learned helplessness, explanatory style, depression, and illness were delayed. And yet, no media scandal resulted, as likely would have been the case in the medical literature. No discussion of the ethics of statistical inference ensued, as also would have likely been the case in the medical literature.

Why Medicine and Psychology Approach Statistical Reform Differently

Why is there such a stark contrast in the way the two disciplines have dealt with the misapplication of NHST? The following section outlines three hypotheses that may explain the difference between medicine and psychology. I stress that these hypotheses may explain the difference in responsiveness to statistical reform in these two disciplines. To put it another way, belief in these three hypotheses is perhaps sufficiently widespread to have impeded statistical reform in psychology. I am *not* arguing that there is a difference in the need for ethical statistical practice in the two disciplines. On the contrary, my aim is to demonstrate that ethical practice is equally important in both.

1. *The proximity of experimental outcomes to utilitarian consequences.* In medicine, trials are usually designed with some particular utilitarian outcome in mind—that is, to test whether a certain specific intervention improves health care in some particular way. By contrast, many psychological trials are designed with the purpose

of improving our understanding of how the mind works, rather than whether one particular intervention improves its function. There are utilitarian ethical arguments to be made here, too, of course, but the ethical consequences of "our theory is wrong" are considerably different than the ethical implications of "our treatment doesn't work" or "our drug causes harm." (There are exceptions in both disciplines, of course.)

2. *The stakes.* In medicine, the stakes can be life or death. Perhaps more often medical studies offer opportunities to enhance health care—for example, through injury prevention, minor improvements to quality of life, and decreases in hospital visits or length of stay. Although not as high, these stakes are still tangible, and medical trials are directed toward measuring these precise outcomes. In psychology, too, there may well be high stakes—opportunities to implement clinical, developmental, or educational interventions, and studies with implications for legal decisions in areas such as child custody or employment—but these outcomes are usually several steps removed from the health results measured by medical trials and experiments.
3. In medicine, distrust of pharmaceutical companies and their motives has led to increased vigilance and helped create a healthy skepticism. In psychology, there is rarely a big, special interest company to blame or substantial potential financial gain by pursuing fallacious results. As a result, it is perhaps more difficult to get a handle on why the statistical errors in psychology should be conceptualized as ethical, as well as academic, concerns.

Costs to Psychology of Not Ethically Motivating Statistical Reform

Whatever the reason for the difference, it is difficult to deny that psychology would be better off if statistical inference was an ethical concern and not just a technical one. There are *costs to science*, as well as costs beyond science, in making ongoing resistance to statistical reform an increasing ethical concern. Below I list some of these costs; there are no doubt others I have not listed.

Costs to science of overreliance on NHST, and neglect of statistical power and MA:

- Research programs may go astray while attempting to explain illusory variability in results (e.g., the employment testing case)

- Unnecessary, prolonged debate; delayed progress and implementation of interventions (e.g., the learned helpless case)
- Time and resources wasted on incorrect, weak, or trivial research programs that happen to turn up statistically significant results
- Potentially useful research programs or directions come to an end because of the inability to produce "consistent" (e.g., statistically significant) results
- Alternatively, research programs may never get started because of their inability to jump the statistical significance hurdle

Costs to public welfare beyond scientific knowledge itself:

- The delayed release of useful interventions and applications (e.g., those based on understanding links between illness, depression, and learned helplessness)
- The implementation of interventions that have little or no impact (in cases where statistical significance has been achieved by overpowered experiments)
- The implementation of harmful interventions because the adverse effects are not detected in low-powered studies
- The implementation of interventions based on misguided theory (e.g., workplace-specific employment tests based on the TSS)
- Various economic costs, including running extra studies in search of statistical clarification when research resources could be better used elsewhere

Conclusion

Thus far, I have discussed the swifter improvements in statistical practice that accompanied the use of an ethical imperative in the medical field. This discussion can be consolidated into the following three main lessons for psychology, from medicine.

Lesson 1: Statistical Reform Needs to Be Ethically, and Technically and Philosophically Motivated

Despite struggling with reform debates for an extra 2 decades, psychology still relies almost exclusively on NHST. Repetition of the technical and

philosophical arguments has done little to motivate change, but psychological researchers may well respond to ethical arguments for statistical reform, as did medical professionals.

Lesson 2: Changes in Statistical Reporting Are Just the First Step; Thinking and Interpretation Need to Change, Too

Thus far I have argued that medicine has improved its practice by treating statistical practice as an ethical issue. I now turn to the more subtle distinction between statistical reporting practice (which medicine has improved dramatically) and *statistical thinking* and interpretation (which has changed far less). Reporting practices in medical journals have changed to be sure, but here is a lesson that psychology can learn from what medicine did not do! Despite the dramatic increase in CI reporting in medical journals, there was little change the way researchers interpret and discuss their findings.

In our own survey of medical journals, we found many articles that did not include any *p* values but still had discussions focused on "statistical significance" (Fidler et al., 2004). Savitz, Tolo, and Poole (1994) also found this in their survey of 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). Poole lamented the fact that change in reporting had not led to a change in thinking:

> The reporting of confidence intervals really hasn't changed the way people think: 99% of the people that now report CIs, 20 years ago would have reported *p* values or asterisks, or *s* and *ns* and they aren't thinking differently to that. They have this vague idea that they are reporting more information with CIs, because they read that somewhere in something Ken Rothman [who made editorial changes at *AJPH* and *Epidemiology*] wrote. But basically they are only reporting CIs because Rothman was an authority figure, and he and others encouraged them—well, his journals insisted on it. (Poole, personal communication, September 2001)

Psychology has the chance to make a substantial reform, one that involves changes in the way researchers approach analysis, and interpret and think about data, as well as what they report in the tables and figures of their journal articles. Substantial reform requires cognitive change and empirical evidence about which presentations of statistics communicate most clearly. Psychology itself is perfectly positioned to collect such empirical data through research into statistical cognition and to advocate for evidence-based reform.

Lesson 3: Statistical Reform Should Be Integrated: Estimation and *Cumulative Meta-Analysis* Go Hand in Hand

Importantly, all four of the case studies given above feature MA. It is one of the best tools available for telling us when there is enough research on a topic for us to stop throwing resources at it. MAs, with their superior power and ability to pin down effect sizes, can help emphasize thinking in terms of estimation rather than hypothesis testing—they stop us from falling into the trap of trying to "explain" the difference between "inconsistent" results of NHST studies. In medicine, editorial policies instituting CIs were a phenomenon of the 1980s, whereas policies about cumulative research and MA came decades later (e.g., 2005 in *The Lancet*). In psychology we should aim to make the shift to CIs and estimation inseparable from cumulative MA: CIs and estimation should be reported as the primary outcome of individual studies, and cumulative MAs should be updated with each new study. The ethical advantages of the two practices in combination are great: CIs make the uncertainty of each trial explicit, and MAs help guard against unnecessary further studies when a question has been adequately answered.[3]

References

Altman, D. G. (1982a). How large is a sample? In D. G. Altman & S. M. Gore (Eds.), *Statistics in Practice* (pp. 6–8). London: BMJ Books.

Altman, D. G. (1982b). Misuse of statistics is unethical. In D. G. Altman & S. M. Gore (Eds.), *Statistics in Practice* (pp. 1–2). London: BMJ Books.

Altman, D. G. (1994). The scandal of poor medical research. *British Medical Journal, 308,* 283–284.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Journal of the American Medical Association, 268,* 240–248.

Chalmers, I. (2005). The scandalous failure of science to cumulate evidence scientifically. *Clinical Trials, 2,* 229–231.

Chalmers, I. (2006). Archie Cochrane (1909–1988). The James Lind library. Retrieved from http://www.jameslindlibrary.org

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*, 12–37.

[3] Author note: This research was supported by the Australian Research Council.

Chalmers, I., Enkin, M., & Keirse, M. J. N. C. (Eds.). (1989). *Effective care in pregnancy and childbirth*. Oxford, UK: Oxford University Press.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services.* London: Nuffield Provincial Hospitals Trust.

Cochrane, A. L. (1979). 1931–1971: A critical review, with particular reference to the medical profession. In G. Feeling-Smith & N. Wells (Eds.), *Medicines for the year 2000*. London: Office of Health Economics.

Cochrane, A. L. (1989). Foreword. In I. Chalmers, M. Enkin, & M. J. N. C. Keirse (Eds.), *Effective care in pregnancy and childbirth*. Oxford, UK: Oxford University Press.

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., … Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

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.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology*. (Unpublished doctoral dissertation). University of Melbourne, Australia.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., … 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 can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15,* 119–126.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gilbert, R., Salanti, G., Harden, M., & See, S. (2005). Infant sleeping position and the sudden infant death syndrome: Systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International Journal of Epidemiology, 34,* 874–887.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1–20.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed). Thousand Oaks, CA: Sage.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine, 327,* 248–254.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

May, W. W. (1975). The composition and function of ethical committees. *Journal of Medical Ethics, 1,* 23–29.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Moore, T. (1995). *Deadly medicine*. New York: Simon and Schuster.

Newell, D. J. (1978). Type II errors and ethics. *British Medical Journal, 5,* 534–535.

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: J. Wiley & Sons, Inc.

Robins, C. J. (1988). Attributions and depression: Why is the literature so inconsistent? *Journal of Personality and Social Psychology, 54,* 880–889.

Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.

Savitz, D. A., Tolo, K., & Poole, C. (1994). Statistical significance testing in the *American Journal of Epidemiology*, 1970–1990. *American Journal of Epidemiology, 139,* 1047–1052.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–315.

Seligman, M. E. P. (1975). *Helplessness: On depression, development, and death*. New York: W. H. Freeman.

Seligman, M. E. P. (1990). *Learned optimism.* New York: Knopf.

Seligman, M. E. P., Maier, S. F., & Geer, J. (1968). The alleviation of learned helplessness in dogs. *Journal of Abnormal Psychology, 73,* 256–262.

Sweeney, P. D., Anderson, K., & Bailey, S. (1986). Attributional style in depression: A meta-analytic review. *Journal of Personality and Social Psychology, 50,* 974–991.

Teo, K. K., Yusif, S., & Furberg, C. F. (1993). Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction: An overview of results from randomized controlled trials. *Journal of the American Medical Association, 270,* 1589–1595.

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

Young, C., & Horton, R. (2005). Health module page. *The Lancet, 366*, 107.