FROM STATISTICAL SIGNIFICANCE TO EFFECT ESTIMATION:

STATISTICAL REFORM IN PSYCHOLOGY, MEDICINE AND ECOLOGY

Fiona Fidler

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy

November 2005

Department of History and Philosophy of Science
The University of Melbourne

ABSTRACT

Compelling criticisms of statistical significance testing (or Null Hypothesis Significance Testing, NHST) can be found in virtually all areas of the social and life sciences—including economics, sociology, ecology, biology, education and psychology. Because it is the overwhelmingly dominant statistical method in these sciences, criticisms need to be taken seriously. Yet, after half a century of cogent arguments against NHST and calls to adopt alternative practices some disciplines, such as psychology, show little sign of change. One obvious question is 'why?' Why are psychological researchers so unwilling to abandon this flawed practice? In this thesis I attempt to answer this question, and compare their practice with other disciplines.

In medicine, effect estimation (in the form of confidence intervals, CIs) was institutionalised in the 1980s through strict and enforced journal editorial policy. It was facilitated by the timely rewriting of textbooks and statistics curricula. The transition was perhaps straight-forward, given the interaction between medical researchers and statisticians, and the processes of statistical editing and reviewing in the discipline. Whilst medicine remains far from a perfect paradigm of statistical practice, it has, on this narrow criterion—deemphasising statistical significance in favour of effect estimation—progressed further than psychology. Ecology too seems to have made some recent progress, though reform remains in nascent stages.

In the absence of adequate guidance from institutions such as the American Psychological Association, and the absence of appropriate editorial pressure, statistical reform in psychology has an uncertain future. What is lacking in psychology (and other disciplines) is an evidence base for statistical reform. This will entail providing empirical justification for adopting alternatives to NHST, and evidence-based guidance for implementing and interpreting those alternatives. The preliminary empirical work in this thesis suggests that CIs do indeed have the necessary cognitive advantage.

DECLARATION

This is to certify that

(i) the thesis comprises only my original work towards the PhD except where indicated in the Introduction;

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

……

Fiona Fidler

PREFACE

During the time I have been writing this PhD, I have had the opportunity to work on two relevant Australian Research Council funded projects—one on statistical reform in psychology, the other on statistical reform in ecology—which have resulted in a number of joint publications. In the chapter outline that follows I explain which sections of the thesis form those publications.

Chapter One documents the uptake of NHST in the three disciplines, exploring in particular the early and strong attachment psychology had to the methods. Results of the journal survey presented in this chapter, and some of the surrounding discussion, has been published as Fidler, Cumming, Burgman and Thomason (2004).

Chapter Two has two distinct parts. First, I catalogue the many criticisms that have been made of NHST over the last half a century, in several disciplines. None of these criticisms are original; they have all been documented before. This part of the chapter, then, is merely a literature review. Second, I review and evaluate defences of NHST that have appeared relatively recently in psychological literature.

In Chapter Three, I argue that typical NHST practices have damaged the progress of these three sciences. As evidence, I provide a series of case studies of particular research programs in psychology, medicine and ecology that have been led astray or otherwise disrupted by problems associated with NHST.

In Chapter Four I turn specifically to statistical reform in psychology. Here I provide a history of particular events—published criticisms, editorial and institutional interventions—aimed at improving statistical practice in psychology, and evaluate the impact of these events. Part of this chapter, specifically the survey of the effects of Philip Kendall's editorial intervention at the *Journal of Consulting and Clinical Psychology*, has been published. The results and a more extensive discussion of the findings can be found in Fidler, Cumming, Thomason et al. (2005). In this chapter, I demonstrate that to date there has been limited response to reformers' calls for a change in psychology. I also introduce an extensive series of interviews with advocates of reform and members of the APA Board of Scientific Affairs' Task Force on Statistical Inference (TFSI), which provide material for several subsequent chapters. (A full list of interviewees and correspondents follows my Acknowledgements.)

Chapter Five critically evaluates attempts to reform the statistical guidelines in the fifth edition of the APA *Publication Manual* (2001), pointing to many reasons why new recommendations are likely to be unsuccessful at motivating change. This chapter is a very slightly modified version of Fidler (2002) and relies heavily on the interviews I described above.

Chapter Six chronicles reform events in medicine (as Chapter Four did for psychology), demonstrating a dramatic shift from NHST to CIs in the mid 1980s. One section of this chapter reports a survey of Ken Rothman's editorial reforms at *The American Journal of Public Health* and *Epidemiology*. This survey and some of the surrounding discussion was published as Fidler, Thomason, Cumming, Finch and Leeman (2004).

Chapter Seven compares reform in psychology and reform in medicine and asks why medicine was able to institute changes to reporting practices when psychology has largely failed to. It also acknowledges that medicine is far from a perfect paradigm of practice, and that both disciplines have some way to go. Some of the discussion here was also published in Fidler, Cumming, Burgman and Thomason (2004).

The focus of Chapter Eight is Ecology. Reform in ecology has progressed in a reasonably different fashion to either psychology or medicine. It is far more focused on Bayesian and information theoretic methods than effect sizes and CIs. Results from the journal survey presented, and surrounding discussion, have been accepted for publication in *Conservation Biology* (Fidler, Burgman, Cumming, Buttrose & Thomason, 2005).

Chapter Nine introduces the notion of evidence-based statistical reform, based on statistical cognition research. As I have explained, an estimation approach has been advocated (as a supplement or replacement to NHST) in many sciences for decades. Yet, few empirical questions have been asked about whether this new approach will be better understood, alleviate widespread misconceptions or lead to more substantial interpretations of research findings. Chapters Nine and Chapter Ten present my preliminary empirical efforts to establish such a research program.

Results from studies presented in Chapter Nine provide empirical evidence that CIs help alleviate a particularly serious misconception associated with NHST, namely that statistical non-significance is equivalent to evidence of 'no effect'. However, there is less positive news in Chapter Ten. Studies in Chapter Ten reveal that CIs themselves are prone to a new set of unexpected and uncontroversial misconceptions. It is too early

to tell whether this is simply because CIs are unfamiliar. Perhaps with adequate training, better presentations and appropriate guidelines, such misconceptions would simply disappear? Whilst this question remains largely unanswered, these two chapters highlight the remarkable fact that, to date, statistical reform has been advocated—and in some disciplines even instituted—without an evidence base. One of the most compelling arguments against NHST is its tendency to be misinterpreted. If it is to be abandoned largely because of this, then surely the onus is on us to provide some evidence that whatever replaces is will be less frequently misunderstood. I conclude this thesis with some brief thoughts about future research directions.

Note: For empirical studies reported in this thesis I have calculated 95% CIs for proportions. Some CIs have been calculated according the method recommended for proportions by Newcombe and Altman (2000); others have been calculate using the standard non-corrected formula for CIs for proportions (z*√(p*(1-p)). Those calculated using the standard formula are therefore prone to the problems Newcombe and Altman suggest. These CIs were calculated and published before I became aware of Newcombe and Altman's recommendations. I have chosen to leave them in their original form so that they correspond with already published figures. In cases where I have used Newcombe and Altman's method (for work already published using this method or unpublished material) this is indicated in text, tables and/or figure captions.

A Further Note: Although the style (referencing, heading, figures and tables) used in this thesis is heavily modelled on APA style, I have sometimes deviated from APA recommendations. I am aware of this, and remind the reader that this thesis is technically not a psychology thesis but rather a History and Philosophy of Science thesis. I have, of course, made every effort to at least be consistent in style.

A final Note: There are several conventions for notation relating to *p* values. I have used '*p* value', but others use 'p-value' or 'P value'. In quotations throughout this thesis, the author's original notation is used.

ACKNOWLEDGEMENTS

During the course of my research I interviewed many *extraordinary* people. Each gave generously of their time and told me almost all of the interesting things I know. A list of interviewees follows the introduction to this thesis. There are also many others, across several departments and universities, who deserve thanks.

Department of History and Philosophy of Science (HPS), University of Melbourne:
Neil Thomason was my primary thesis supervisor. I have never met anyone like Neil. During the years we have worked together he has been relentlessly curious, enthusiastic and extremely generous. He has also been kind and patient. Amongst so many other things, Neil taught me how to be a good teacher.

Keith Hutchison co-supervised this thesis (in an official capacity); he signed a lot of forms and changed the way I give conference papers—for the better! Thanks also to Rosemary Robins and Howard Sankey.

The HPS postgraduate association (circa 1999-2003):
Marg Ayre, who taught me what 'finishing' meant, years before I could understand; Tao Bak; Emmaline Bexley, who, amongst so many other things, helped me format this thing; Kristian Camillieri; David Evans; Matthew Klugman, who read several draft chapters of this thesis and taught me much about writing history—remaining flaws of technique are of course, my own; Les Kneebone (whom I will thank again); Claire Leslie, Erik Nyberg and Peter Parbery.

Environmental Science Lab (and related parties), School of Botany, University of Melbourne:
A large lab, packed with very kind and very smart people. Marg Burgman let me work there for over four years—probably much longer than he expected. Amongst others I worked with: Joe Banks; Sarah Bekessy (who helped make my survey questions to ecology students plausible); Jan Carey (thanks for debates about power and confidence intervals, and for chocolate); Yung En Chee (thanks for understanding and see below); Ryan Chisholm; Jane Elith (who has saved me from computer diasters many times and in so many ways, takes care of us all); Michelle Ensbey; Frith Jarrad (who may have

suffered the most, as a consequence of lab geography); Lauren Keim (who read a chapter and was nice); Prema Lucas (who didn't but is forgiven); Mick McCarthy (who convinced me that Bayes made sense); Kirsten Parris (for sharing stories about frogs and editors); Cassia Reid; Tracey Regan (who showed me what it was all about); Andrea White (whose lunch packs have kept me alive the last few months); Bonnie Wintle and Brendan Wintle (who very patiently explained AIC and other important things).

School of Psychological Science, La Trobe University:
Where I spend a lot of my time now: Cathy Faulkner read chapter drafts, offered invaluable advice, was patient and listened a lot—the clinical psychologist every PhD candidate should know. I look forward to continued collaborations! Thanks also to Melissa Coulson; Sarah Belia; Jo Leeman and my third year research methods students.

A particular acknowledgement under this heading to Geoff Cumming—an ever watchful 'step-supervisor' who for some reason, years ago, put his faith in a strange HPS student and continues to offer seemingly endless opportunities to extend projects. Totally unimaginable that this thesis would have come together without him! For one thing, I wouldn't have understood confidence intervals without his pictures; for another, I may have inappropriately used '/'.

Money:
I am grateful for these scholarships and grants: Melbourne Research Scholarship (MRS); Melbourne Abroad Travelling Scholarship (MATS); Travel Research in Postgraduate Study (TRIPS). In addition, I have worked on projects funded by The Australian Research Council (ARC). Thank you to those who bought home the ARC grants: Mark Burgman, Geoff Cumming, Neil Thomason and Sue Finch.

Family and Friends:
Thanks for putting up with me: Toni Fidler (my mum); Frank Fidler (my dad); Jo Roxburgh; Robert Roxburgh; Eugine and Kay Kneebone; Les Kneebone; Emmaline Bexley; Tracey Regan; Helen Regan (whose hospitality helped 'fund' many an overseas jaunt); Marg Ayre; Yung En Chee; Andy White; Naomi Toottell; Lauren Keim;

Matthew Klugman; Briar Ballantyne; Chris Brent; Darrian Collins; Julie Connolly; Sharna Hackett; Ros Irons; Gabe Kneebone; Nikki Lesley.

Special thanks to Zoe Loh, who on top of everything else, discovered and collaborated with me on 'The Case of Spontaneous Recovery' (Chapter Three). We presented a joint paper: Loh, Z. & Fidler, F. (2001) "Null Hypothesis Significance Testing: Can the damage be assessed?" Paper presented to the *Australasian Association of History, Philosophy and Social Studies of Science (AAHPSSS)*, Melbourne (July 2001).

Shared Spaces:
Strange, and occasionally wonderful, things happen in shared work spaces. These people were in close proximity through the best and/or worst: Sarah Bekessy; Dave Duncan; Cathy Faulkner; Frith Jarrad; Tracey Regan and Andy White.

Also:Yung En Chee (technically not office-sharing, but I spent so much time asking for her help…); Emmaline Bexley (same technicality, but our shared house often became an office, except with better food and wine) and Les Kneebone (for the same reason).

Specialist Librarian: Les Kneebone

Where does Sue Finch fit in? So many places! For one, the *Department of Mathematics and Statistics, University of Melbourne*. Thank you, Sue, for so many things including: initial formal supervision (through the Mathematics and Statistics Department); technical and survey design advice; 'management' advice; holidays in the Netherlands; ongoing collaboration and, of course, reading this thesis. Also thank you for coffee and for always sending Christmas cards.

Finally, thank you to the examiners of this thesis for their valuable suggestions.

LIST OF INTERVIEWS

Psychology

Mark Appelbaum:
- Department of Psychology, University of California, San Diego, USA; APA Task Force on Statistical Inference (TFSI) member; past-president of APA Division 5 (mathematical and statistical interest group); past editor of *Psychological Methods*.

Patricia Cohen:
- New York State Psychiatric Institute, Columbia University, New York, USA; wife and collaborator of the late Jacob Cohen.

Geoff Cumming:
- School of Psychological Science, La Trobe University, Melbourne, Australia..

David Grayson:
- School of Psychology; University of Sydney, Australia.

Janet (Shibley) Hyde:
- Department of Psychology, University of Wisconsin, Madison, USA; past chair of the Publications and Communications committee of the APA.

Daniel Kahneman:
- Department of Psychology, Princeton University, USA; Recipient of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, 2002.

Roger Kirk:
- Department of Psychology and Neuroscience, Baylor University, Waco, USA; past-president of APA Division 5.

David Krantz:
- Department of Psychology, Columbia University, New York, USA.

Geoff Loftus:
- Department of Psychology, University of Washington, Seattle, USA; past editor of *Memory and Cognition*.

Paul Meehl (1920-2003):
- then at: Department of Psychology, University of Minnesota, Minneapolis, USA; expert advisor to the APA TFSI; past president (1962) of the APA.

Sangeeta Panicker:
- APA Director of Research Ethics, Washington, DC; APA staff liaison to the TFSI.

Pip Patterson:
- Department of Psychology, University of Melbourne, Australia.

Robert Rosenthal:

- Department of Psychology, University of California Riverside, USA; co-chair of APA TFSI.

Joe Rossi:

- Cancer Prevention Research Center, University of Rhode Island, USA.

Liora (Pedhazur) Schmelkin:

- Vice Provost for Academic Affairs, Hofstra University, Hempstead, USA; past-president of Division 5.

Patrick Shrout:

- Department of Psychology, New York University, USA.

Neil Thomason:

- Department of History and Philosophy of Science, University of Melbourne, Australia.

Bruce Thompson:

- Department of Educational Psychology, Texas A&M University, USA; member of APA TFSI; past editor of *Educational and Psychological Methods.*

Others imparted crucial knowledge, but I was unable to tape our discussions (usually due to meeting circumstances rather than declines to be taped):

Leona Aiken:

- Department of Psychology, Arizona State University, USA; member of APA TFSI.

Gerd Gigerenzer:

- Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.

Lisa Harlow:

- Cancer Prevention Research Center, University of Rhode Island, USA.

Frank Schmidt:

- Department of Management and Organizations, University of Iowa, USA.

Others provided correspondence:

Siu Chow:

- Department of Psychology, University of Regina, Canada; defender of NHST.

Richard Harris:

- Department of Psychology, University of New Mexico, USA.

Michael Smithson:

- School of Psychology, Australian National University, Canberra, Australia.

Medicine and Epidemiology

Ken Rothman:

- Vice President of RTI Health Solutions; Professor of Epidemiology and Professor of Medicine at Boston University, USA; past-president of the Society for Epidemiologic Research; Honorary Fellow of the American College of Epidemiology; past assistant editor of *American Journal of Public Health;* founding editor of *Epidemiology*.

Charles Poole:

- School of Public Health, University of North Carolina, Chapel Hill, USA

Others provided correspondence:

Janet Lang

- The Watson Institute for International Studies, Brown University, USA; past co-editor of *Epidemiology*.

Geoff Berry

- School of Public Health, University of Sydney; past editor of *Medical Journal of Australia.*

Ecology

David Anderson:

- Colorado Cooperative Fish and Wildlife Research Unit and Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, CO, USA

Mark Burgman:

- Environmental Science Group, School of Botany, University of Melbourne, Australia.

Mick McCarthy:

- Australian Research Centre for Urban Ecology, School of Botany, University of Melbourne, Australia

Kirsten Parris:

- School of Ecology and Environment, Deakin University, Melbourne, Australia.

Others provided correspondence:

- The environmental science lab at the University of Melbourne

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students (Rozeboom, 1997, p. 335).

This thesis is about the prolonged controversy over Null Hypothesis Significance Testing (NHST) in medicine, ecology and particularly, psychology. The work here is partly historical, partly philosophical and partly sociological. It also includes empirical studies of statistical cognition. In bringing together these traditions of analysis, as well as examining three disciplines, it is perhaps a truly multi-disciplinary work.

It did not, however, start out that way. My original intention was to explore researcher resistance to statistical reform in psychology. My background is in psychology, and I unwittingly uncovered a little piece of the controversy in my final undergraduate year[1]. I was attempting a power analysis, as part of designing a research project for my honours thesis, and encountered some unexpected difficulties. In undergraduate quantitative lectures we had been introduced to the concept of statistical power—it was presented as an important issue, worthy of serious attention. In these lectures, we had been taught how to calculate power in the context of a simple, two independent groups research design. I was now faced with the prospect of a complicated mixed design with several independent and dependent variables.

As I made my way around the offices of quantitative types in the psychology department, I realised I was not the only one broaching these questions for the first time. For almost ever academic whose advice I sought, it was also the first time they had considered the question of statistical power in practice. Most attempted to dissuade me from taking the question further, and dismissed the idea that statistical power was as important as I had come to believe from my lectures. I was, for example, reassured I would not loose grades for failing to report a power analysis. Only one person recommended I visit the department of statistics for help (I did visit, but it didn't help). I never did find out how to do the calculation, but I went ahead with the study anyway.

[1] In Australian universities the fourth year of a Bachelor degree is called an 'honours' year. It consists of some courses and a reasonably large independent research project (written into a final report of roughly 15,000).

The year before I started my honours project, a substantial meta-analysis on the topic I was investigating (alcohol and information processing) had been published in a well known experimental journal[2]. The meta-analysis had identified a general linear effect of alcohol on global information processing, rather than specific effects on particular stages in an information processing task. My study failed to replicate the statistically significant global response in the meta-analysis. Without any consideration for statistical power, I wrote up my single study as though it contradicted the meta-analytic findings. All this may be a shocking admission for an honours student, but the worst is yet to come—my thesis received the best grade in my year and I won the university's Australian Psychological Society Prize. The glaring methodological flaw went completely unnoticed by my peers, teachers and examiners.

My interest in the questions of *this* thesis grew, in part, from those experiences. Of course, it was only years later that I realised the relationship between statistical power analysis and meta-analysis, and how (potentially) foolish the argument in my project report had been. But the curiosity of being taught an important concept (statistical power), and then being discouraged from following through with the analysis in a real research situation, struck me immediately. I also realise now that, as an undergraduate psychology student over a decade ago, I was perhaps lucky to have been exposed to statistical power at all.

When I decided to undertake this PhD topic, I was interested in why researchers in psychology were 'resisting' statistical reform. I had framed the question as one primarily about individual psychological researchers' cognitive processes. It soon became apparent that the story was much bigger than I had anticipated—in two ways.

First, problems with NHST are present in many disciplines, not just psychology. It did not take long to discover that criticisms of NHST had been repeated for decades in almost all areas of the social and life sciences. Second, any explanation that located the failure of reform in the mind or actions of the individual researchers (as in 'researcher resistance') was rapidly exposed as superficial, especially when the other disciplines were taken into account.

Criticisms of NHST and attempts at reform have, of course, a long history. Ecology had its first wave of criticisms of NHST in the 1980s. By this time they were

---

[2] The meta-analytic study was Maylor, E.A. & Rabbitt, P.M.A. (1993). Alcohol, reaction time and memory: A meta-analysis. *British Journal of Psychology, 84,* 301-317.

already well-known in other disciplines. In medicine, criticisms started in the late 1960s (Cutler, Greenhouse, Cornfield & Schneiderman, 1966); and in psychology, earlier still (Abelson, 1954; Meehl, 1954, 1957). Independent criticisms by statisticians date back to at least the 1930s (Berkson, 1938, 1942). Even before this, however, NHST had been controversial. Fisher never agreed with the Neyman-Pearsonian modifications of hypothesis testing and Neyman and Pearson saw Fisher's theory as incomplete. In a sense, NHST was born into controversy.

For psychology, earlier critics did not spell earlier reform. Despite literally hundreds of psychology articles criticising the practice, and various editorial efforts, including those initiated by the American Psychology Association (APA), reporting of *p* vales and statistical significance still dominates the literature, as many journal surveys demonstrate. Medicine, on the other hand, was more successful in moving from statistical significance to estimation (i.e., effect sizes and precision in the form of confidence intervals, CIs). In medicine, reporting practices underwent considerable reform in the mid 1980s—around 20 years ago! Why was medicine able to institute such change when psychology was not and still seemingly can not? Have events in ecology been more like medicine, or more like psychology? What will its future hold? What can disciplines like psychology do to motivate change now? Are some reforms or changes more effect than others? What evidence justifies various alternatives to NHST? These are some of the more pragmatic (and more recent) questions driving this thesis.

There are essentially two categories of criticisms of NHST. First, there are those that take issue with frequentist philosophy of statistics in general (most notably, Bayesians), and second, there are the majority, that do not. Criticisms of null hypothesis testing from within a frequentist framework are not necessarily limited to those of 'misuse' or 'misinterpretation' of the test (e.g., neglect of statistical power or misinterpretation of a *p* value). They also legitimately address the irrelevance of dichotomous decision-making, neglect of estimation, and the limited emphasis on prior information and cumulative knowledge. Many of these concerns are shared by Bayesians, but they are not intrinsically Bayesian.

In this thesis I briefly cover broader criticisms, such as Bayesian criticisms, but my focus is largely elsewhere. I am not neglecting Bayesian arguments because of any strong faith in classical statistics but rather because this thesis is about why many researchers and editors have not answered reformers' calls for change—and these calls, overwhelmingly, have come from within a classical statistics framework.

Most calls for reform are modest. They do not require a paradigm shift in probability theory. The most often proposed alternatives would seem to require only relatively minor changes to practice, that is, reporting effect sizes and CIs instead of *p* values or dichotomous accept reject decisions alone. It is the difficulty (in some cases, the virtually impossibility!) of achieving these small changes that makes the NHST controversy such a fascinating philosophy and sociology of science case study. It would perhaps not come as a surprise if a proposed shift from a Frequentist to Bayesian approach took half a century. But in some disciplines, such as psychology, even a comparatively minor shift to estimation has taken this long—and has in fact, not gone far.

What is Statistical Reform?

What should we do instead of significance tests? How do we test hypotheses without significance tests? In my mind, there is nothing to be replaced. If we had a cancerous tumor growing in our lungs, we would ask a doctor to remove the growth; we would not ask the doctor to replace it with a better cancer.
(Richard Fraley, http://www.uic.edu/classes/psych/psych548/fraley/NHSTsummary.htm, accessed 02-02-05).

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

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

some attention—admittedly less than in medicine or psychology—in the ecology reform literature.

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

Yet another recommendation, common in ecology, is the use of information theoretic approaches—a move led by Ken Burnham, David Anderson and colleagues. Akaike Information Criteria (AIC), based on the work of 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 is used to compare competing models, and to combine, or average, models to make multi-model inferences.

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

Whilst all or any of the practices listed above (effect estimation, CIs, standardised effect sizes, information theoretic and Bayesian methods) constitute statistical reform, so do more generic practices, such as: increased use of graphical representations; consideration of the clinical or biological or practical importance of results (as opposed to merely the statistical significance); consideration of sample size issues; appreciation of meta-analytic issues and more generally, the need for scientific results to be cumulative; and other thoughtful treatments of trends, patterns and effects that go beyond the mechanised dichotomous decision process of NHST. As I have mentioned my main focus in this thesis is on estimation and precision (i.e., effect sizes and CIs) as an alternative to NHST: why even minimal steps towards going beyond NHST have been so difficult to achieve, and in many cases, are still a way off.

1
NHST AND THE INFERENCE REVOLUTION

They [statisticians] have already overrun every branch of science with the rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle. (Kendall, 1942, p.69).

Null Hypothesis Significance Testing (NHST) was at the forefront of the inference revolution in the social and life sciences between 1940 and 1960. Many statistical significance testing techniques had been developed decades before (e.g., William Gosset published his paper on Student's *t* test in 1908; Fisher published on ANOVA in 1918), but it was largely post World War II that these techniques entered and reshaped these sciences. This chapter explores the rise of NHST in three disciplines—medicine, psychology and ecology.

The subject of statistical significance testing would be unlikely to sustain a thesis of this length if it were not for the astonishing amount of controversy surrounding its use. The criticisms are by no means limited to the three disciplines I focus on here. Similar arguments, made over several decades, can be found with ease in: sociology (e.g., Morrison & Henkel, 1969; Weinbach, 1989); education (e.g., Schafer, 1993; Thompson, 1996); criminology (e.g,. Weisburd, Lum & Yang, 2003); economics (e.g., Altman, 2004; McCloskey, 1992, 1995; Zilak & McCloskey, 2004); marketing (e.g., Sawyer & Peter, 1983); chemistry (e.g., Harris, 1993; Henderson, 1993); nursing (e.g., Glaser, 1996) and, notably, statistics itself (e.g., Hall & Selinger, 1986; Kempthorne, 1976; Royal, 1986).

1.1 Origins of Null Hypothesis Significance Testing

Several secondary sources (e.g., Huberty & Pike, 1999; Hald, 1990; Gigerenzer & Murray, 1987; Stigler, 1999, 1986; Hacking, 1965) credit John Arburthnot's 1710 report on the distribution of male and female births as the first use of a hypothesis test[3]. Arbuthnot collected 82 years of christening records in London to determine whether the birth rate of both sexes was equal. Assuming all babies born were christened,

[3] Arbuthnot's reasoning is most often compared to Fisherian hypothesis testing. However Sober (2005) argued that Arburthnot's approach may have more closely paralleled a likelihood approach than a modus tollens based hypothesis testing approach.

Arbuthnot discovered a slightly higher rate of male births every year. The probability of the higher male birth rate being due to chance, according to Arbuthnot's calculations, was extremely small (i.e., 0.5^{82}). That an event of such low likelihood had in fact occurred, he interpreted as evidence of design in the universe. His reasoning was that young men died routinely in wars and a higher male birth-rate compensated for this maintaining equal numbers of the sexes. Equal number of sexes allowed for monogamous marriage: God's intention, believed Arbuthnot.

Over the next century, understanding of probability distributions grew dramatically with the work Bernoulli, Laplace and others (Huberty & Pike, 1999). There is debate amongst historians of science as to exactly when the era of modern statistical testing began. Formal theories of variation and correlation were not quite present in Charles Darwin's work, though the relevant questions were laid out as though ready for analysis (Cowles, 1989). Some attribute the revolution to Darwin's cousin, Sir Francis Galton, which dates it to the 1870s: "And it is with Galton, who first formulated the method of *correlation* that the common statistical procedures of modern social science began." (Cowles, p.2). Others (e.g., Salisburg, 2001) argue that Galton's disciple, Karl Pearson, deserves the credit. K. Pearson refined Galton's correlation analysis, and developed a goodness-of-fit test based on a chi-square distribution. In any case, it is clear that the birth place of modern statistical tests was evolutionary theory and eugenics, as much as mathematics.

The language of 'statistical significance' appears to have started with Francis Edgeworth. Stigler (1986) reported that it was in 1885 that Edgeworth first coined the term 'significant' to describe a difference between groups. Edgeworth described the spread of a distribution in terms of the 'modulus'— roughly equivalent to $\sqrt{2}$ standard deviations. A difference was considered 'not accidental' or 'significant' if it exceeded two or three times the modulus (Spiegelhalter, 2004).

There are many other developments that might also be considered necessary prerequisites to modern null hypothesis testing—too many to mention. Those that cannot go without mention, however, are the early 20th century developments of William Gosset, Ronald Fisher and Jerzy Neyman and Egon Pearson.

Gosset, Fisher and Neyman-Pearson

In 1908 William Gosset published an article in *Biometrika*. The article contained the first ever *t* distribution. At the Guinness brewery, where Gosset worked,

obtaining a single observation of the relationship between malt and hops could take a whole day and was expensive. Being able to work effectively with small samples had obvious value. Gosset noticed that when sample sizes were very small, the sample standard deviation was an erratic estimate of the population standard deviation. This led him to small sample estimation by means of the *t* distribution. Gosset published under the pseudonym of 'Student'[4] and his test is usually referred to as Student's *t.*

It was not an accident that Gosset's article appeared in *Biometrika*. The journal was largely funded by Francis Galton and it was edited by Karl Pearson. Gosset had previously consulted with K. Pearson and spent some time in 1906 and 1907 working in his lab on sampling distributions (E. Pearson, 1990). *Biometrika* later became a major outlet for the work of Ronald Fisher and the stage for the some of the great early debates over modern statistical theory.

A decade later, in 1918, Fisher published the concept of Analysis of Variance (ANOVA) in a population genetics paper. He wrote: "It is ... desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance..." (p. 399). However, this was not his formal introduction of ANOVA. The formal introduction, and the term 'Analysis of Variance', came a few years later (Fisher, 1921) in the first of his famous series of papers on crop rotation. Fisher, like K. Pearson before him, worked only with a single hypothesis—the null hypothesis. His ANOVA procedure calculated an associated probability, that is, the probability of an observed result, or one more extreme, given the truth of the null. This, of course, is the very familiar concept of the *p* value, and the basis of modern statistical significance test that dominates so many sciences today. Fisher introduced the broader, applied research community to his theory of null hypothesis testing in *Statistical Methods for Research Workers* (1925) and later in *The Design of Experiments* (1935).

Through the mid to late 1920s Jerzy Neyman and Egon Pearson collaborated to further develop what they believed to be Fisher's incomplete theory of statistical testing. To this end, they introduced the concepts of the alternative hypothesis, type I and II errors and statistical power. In 1928 they published two articles in *Biometrika* outlining

---

[4] I have heard two theories about why Gosset's employer, the Guinness brewery, required publication under a pseudonym. One is that upon realising Guinness had hired a statistician, other breweries might follow suit and also hire statisticians and Guinness would loose their competitive edge. The second is that by making public their need for a statistician Guinness would be acknowledging that there was variation in their product and exposing the fact that quality fluctuated.

their own theory of hypothesis testing (Neyman & Pearson, 1928a, 1928b). Then, in 1933, they published their most influential paper "On the problem of the Most Efficient Tests of Statistical Hypotheses" (Neyman & Pearson, 1933).

Both Fisher's approach and that of Neyman and Pearson share a relative frequency theory of probability—meaning probabilities are understood as rates in an infinite long run of trials and not as relating to the occurrence of any single event. However, their interpretations of the theory are quite different. For Fisher, the associated probability (*p* value) was more than a frequency statement; it could also be understood as a degree of belief, or more specifically, a "rational measure of disbelief" (Oakes, 1986, p.5). Fisher's discussion of probabilities as 'measures of belief' does not sit comfortably within the framework of frequentist statistics and some have suggested it has Bayesian overtones (Savage, 1961; Gigerenzer, 1993).

Neyman and Pearson took what is generally considered a strict approach to frequency, disallowing statements about the probability of any single uncertain event. Under this approach, one does not make inferences about individual null hypotheses, but rather decisions about how to act or behave. Inferences are restricted exclusively to events that can be understood as instances of an infinite long run of repetitions under the same conditions. Then notion of probability as degrees of belief is not tolerated at all by this school.

Perhaps the most well-known disagreement between Fisher and Neyman and Pearson—and that for which Fisher has been most often criticised—was over the alternative hypothesis. The Neyman-Pearson approach entails specification of both a null and a specific alternative hypothesis; Fisher recognised only a null hypothesis. Denying the alternative hypotheses of course means denying the measurement of type II errors and the related concept of statistical power. Fisher justified this on the grounds that the purpose of any experiment was only to "give the facts a chance of disproving the null hypothesis" (1935, p.19)—not to accept the null or provide evidence for any alternative hypothesis. The null is rejected when the observed experimental effect falls within the tail areas of a distribution for that parameter. Fisher suggested, on the grounds of convenience, .05 and .01 as cut offs for the tail area. Events at this point or further into the tail had an "exceptionally rare chance" of occurrence if the proposed null was in fact true (Fisher, 1973, p.42). But as Oakes explained: "without reference to an alternative class of hypotheses there is no apparent reason to choose the tail area as a

region of rejection. Any part of the sampling distribution with area of .05 would serve a similar purpose" (1986, p.122).

The term 'significance level' also meant quite different things in each of the two approaches. To Neyman and Pearson it was an *a priori* decision criterion; to Fisher a property of the collected data. For Neyman and Pearson, the exact *p* value provided no further direct evidence for or against the null hypothesis. If, for a particular case, the rejection region α was set at .05, a result of p=.0003 must lead to the same decision as a result of p=.03. Although Fisher also used rejection regions, the exact *p* value was meaningful for him because of his 'rational measure of disbelief' interpretation of associated probability.

And so NHST was born into controversy. Yet its roots in these conflicting statistical theories are rarely acknowledged in the teaching and application of modern statistical methods. Not surprisingly, this has caused much confusion and many of the misconceptions about NHST outlined in Chapter Two are thought to stem from the failure to acknowledge these conflicts. Gigerenzer (1993) provides a well-known summary of this argument:

> What has become institutionalised as inferential statistics in psychology is not Fisherian statistics. It is an incoherent mishmash of some of Fisher's ideas on the one hand and some of the ideas of Neyman and E.S. Pearson on the other… Fisher, Neyman and Pearson would all have rejected it, though for different reasons. The institutionalised hybrid carries the message that *statistics is statistics is statistics*, that is, that statistics is a single integrated structure that speaks with a single authoritative voice… Students and researchers should be exposed to different approaches (not one) to inductive inference and be trained to use these in a constructive (not mechanical) way. A free market of several good ideas is better than a state monopoly for a single confused idea. (p. 311, italics in original).

1.2 Uptake of NHST in Three Sciences

All the major components of NHST had been developed and published by the mid 1930s, yet its spread to sciences outside statistics took, in some cases, decades. This is not surprising. Major upheavals in scientific thought often take time. Given that

this particular scientific revolution was interrupted by World War II, the uptake of NHST might even be considered a relatively rapid event in the history of science. In the remainder of this chapter I document the rise of NHST in three disciplines: medicine, ecology and psychology.

1.2.1 Method

I surveyed selected 1950 to 1970 issues, in 5 year intervals (i.e., 1950, 1955, 1960, 1965, 1970), of journals in medicine and ecology (listed below). For psychology, such surveys already exist, so data for that discipline is taken from published literature, as explained later. My decision to survey the period 1950 to 1970 for medicine and ecology was based on two pieces of information. First, Gigerenzer and Murray (1987) argued that the inference revolution in psychology occurred between 1940 and 1955. Second, textbooks and expository articles suggest the change in psychology predates corresponding changes in other disciplines. In browsing several 1940 issues of journals in medicine, I found no evidence of substantial NHST use during this decade. Further, my survey results show that major changes in statistical practice in medicine and ecology did indeed occur during the time period surveyed.

I coded only 'contributed' or 'research' articles reporting new data. That is, I did not code qualitative case studies, theoretical or purely methodological articles, letters, editorials or any other non-empirical or non-quantitative articles. Articles sampled depended somewhat on the availability of issues and were chosen to represent a cross section of the representative year.

For each article I recorded only whether or not NHST was reported. I did not record the number of occurrences per article. This was to ensure that the NHST reporting rate was not inflated by one or a few articles reporting multiple tests. Any report of a *p* value or reference to a 5%, 10% or other 'level of significance' was counted as an instance of NHST. Similarly, use of the term 'statistically significant' and use of asterisks in tables and footnotes referencing 'significance' were included as instances. Use of the terms 'significance' or 'significant' alone (i.e., not in reference to asterisks, *p* values or not prefaced by 'statistical') was not counted as an instance of NHST. This may have resulted in some 'misses' in my coding, as it is possible that some uses of 'significance' were in fact referring to formal statistical tests, but these cases were too ambiguous to be legitimately included in the analysis.

For each time period, I calculated the percentage of total empirical articles reporting NHST. Percentages in tables and figures are reported with 95% CIs. CIs were calculated using the method recommended for proportions by Newcombe and Altman (2000).

Medicine

In medicine I coded articles from *British Medical Journal (BMJ)*, *Lancet* and *New England Journal of Medicine (NEJM)*. These journals have international reputations and extremely high impact factors[5]: 7.038, 21.713, 38.570 respectively (ISI Journal Citation Reports for Medicine—General and Internal, 2004). For each time period I coded between 55 and 70 articles (exact *n* given in results), in each of the three journals.

Ecology

Ecology, being a relatively new science, had few journals dating back as far as the 1950s. I chose *Ecology* and the *Journal of Ecology* on the grounds that they were indisputably 'ecology' and because they had high impact factors: 4.104 and 3.390 respectively (ISI Journal Citation Report for Ecology, 2004). If these impact factors seem a little low compared to those of the medical journals surveyed above, bare in mind that in the 2004 ISI Journal Citation Report for Ecology there were only two journals[6] with impact factors above 5—both of which are primarily review journals and neither of which date back far enough for this survey.

The number of articles surveyed at each time period varied according to the number of empirical articles published in each period and was always considerably less than in medicine (exact *n* given in results). Because the number of articles in ecology, unlike medicine, allowed it, I coded *all* available empirical articles for the selected years.

---

[5] An impact factor is an average frequency of citations per article, in a given journal per year. The number of citations in the current year's articles to articles in the previous 2 years is summed. This number is then divided by the total number of articles in the previous two years (Journal Citation Reports, 2004).

[6] *Trends in Ecology and Evolution*, impact factor=12.938, started in 1986 and *Annual Review of Ecology Evolution and Systematics*, impact factor=9.429, started in 1970 (ISI, 2004).

Psychology

The case of psychology is unique amongst these three disciplines. In psychology, unlike medicine or ecology, the uptake of NHST is well-documented (e.g., Hubbard & Ryan, 2000; Lovie, 1979; Rucci & Tweeney, 1980). Results and discussion related to psychology are, therefore, reviews rather than an original survey and argument.

1.2.2 Medicine and Ecology: Survey Results and Discussion

There was a dramatic change in statistical practice in medical and ecology journals during the time period surveyed. Figure 1.1 shows the percentage of articles reporting NHST between 1950 and 1970 in each of the three disciplines. The figure shows two sets of data collected by me, for medicine and ecology, and a third set for psychology, taken from Hubbard and Ryan (2000). Percentages for medicine and ecology are given with 95% CI. For ecology I report 95% CIs even though I sampled all eligible articles in a given year (and so in a sense, a population) because I consider those years to be samples of five year intervals. (The psychology data is presented without CIs, as these were not provided in the original. Note, however, the very large sample size used by Hubbard & Ryan (2000) in the figure caption below).

Medicine

In 1950 and 1955 the average reporting rate of NHST in these three medical journals was less than 20% (18% in 1950; 19% in 1955). By 1960 this percentage had increased to 31% and it continued to steadily rise over the next decade. In 1965 it reached 40% and by 1970 around half (51%) of empirical medical articles were reporting *p* values and/or significance levels.

There were some differences between medical journals, particularly in the earlier time periods. Table 1.1 shows the number and percentage (with 95% CIs) of NHST articles for each journal, at each time period. (Figure 1.2 also displays the same percentages[7].) In *BMJ* NHST appeared in over a quarter (27%) of articles in 1950

[7] I have reported percentages and their associated CIs twice—in tables as well as figures. This is somewhat unconventional, and the apparent redundancy requires explanation. In an effort to make figures more readable, I show only the upper half 95% CI. This gives an indication how much higher than my percentage point estimate the true rate of NHST use may have been at each period. Adding the lower half CI made even these relatively simple figures obscured pattern and made interpretation difficult.

compared to just 13% and 11% in *Lancet* and *NEJM*. However, whilst NHST reporting in *Lancet* and *NEJM* rose fairly consistently after this time, *BMJ* fluctuated—dropping in 1955 to just 16% and then rising again to 43% (again well ahead of the others) in 1960. There is no obvious explanation for this pattern of inter-journal differences nor anything to suggest that it is particularly meaningful or important.

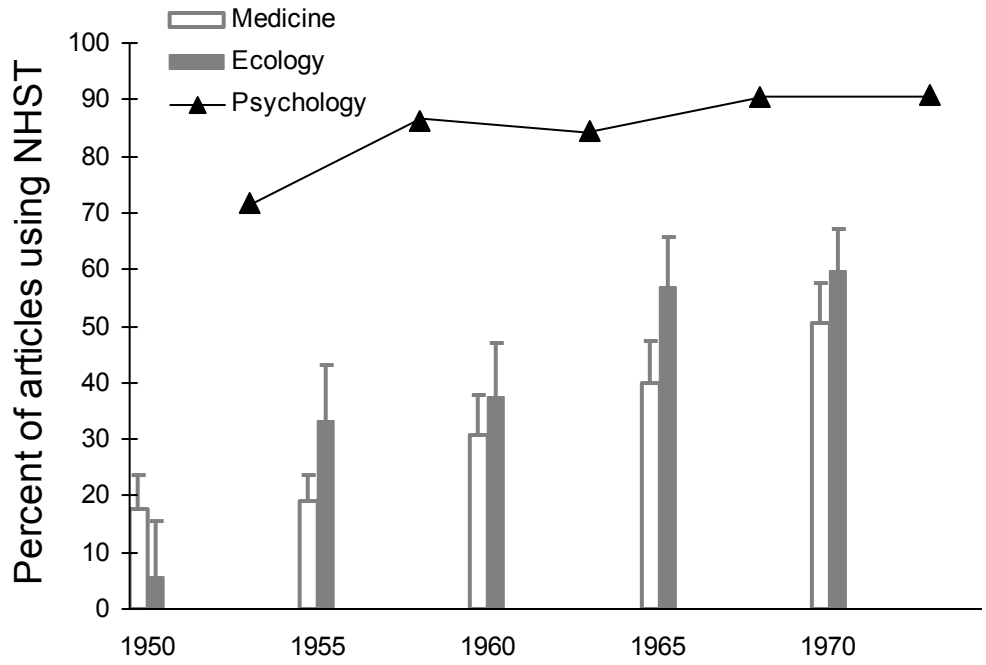


Figure 1.1. Percentage of research articles published between 1950 and 1970 in medicine, ecology and psychology reporting NHST. Medical articles (total n=913) in each of the years graphed from *British Medical Journal*, *Lancet* and *New England Journal of Medicine*. Ecology articles (total n=524) in each of the years graphed from *Ecology* and *Journal of Ecology*. The percentages of NHST reporting in psychology are taken from Hubbard and Ryan's (2000) survey of 12 APA journals[8]. Shown here are Hubbard & Ryan's percentages for 1950-1954, 1955-1959, 1960-1964, 1965-1969 and 1970-1974 (total n=3417). Error bars are upper half 95% CIs based on Newcombe and Altman's (2000) recommended method for proportions.

This obstacle to CI reporting is discussed again in Chapter Ten of this thesis. Yet, figures are important for conveying over-all trends. I have therefore included CIs in tables as well, so that the lower limit is provided for the reader. This is particularly important in the case of proportions, as the appropriate CI method (Newcombe & Altman, 2000) often results in intervals that are not symmetric about the percentage point estimate.

[8] *American Psychologist*, *Developmental Psychology*, *Journal of Abnormal Psychology*, *Journal of Applied Psychology*, *Journal of Comparative Psychology*, *Journal of Consulting and Clinical Psychology*, *Journal of Counseling Psychology*, *Journal of Educational Psychology*, *Journal of Experimental Psychology—General*, *Journal of Personality and Social Psychology*, *Psychological Bulletin* and *Psychological Review*.

Table 1.1.
Percentage of *British Medical Journal*, *Lancet* and *New England Journal of Medicine* articles published 1950 to 1970 reporting NHST.

	BMJ	Lancet	NEJM	Total
1950	27% (18 of 67)	13% (7 of 55)	11% (6 of 55)	18% (31 of 177)
95% CI	18 to 39%	6 to 24%	5 to 22%	13 to 24%
1955	16% (11 of 68)	29% (16 of 55)	14% (8 of 59)	19% (35 of 182)
95% CI	9 to 27%	19 to 42%	7 to 24%	14 to 26%
1960	43% (29 of 67)	25% (15 of 61)	23% (13 of 57)	31% (57 of 185)
95% CI	32 to 55%	16 to 37%	14 to 35%	25 to 38%
1965	47% (33 of 70)	33% (19 of 58)	39% (22 of 57)	40% (74 of 185)
95% CI	36 to 59%	22 to 46%	27 to 52%	33 to 47%
1970	48% (33 of 69)	47% (26 of 55)	57% (34 of 60)	51% (93 of 184)
95% CI	36 to 59%	35 to 60%	44 to 68%	43 to 58%

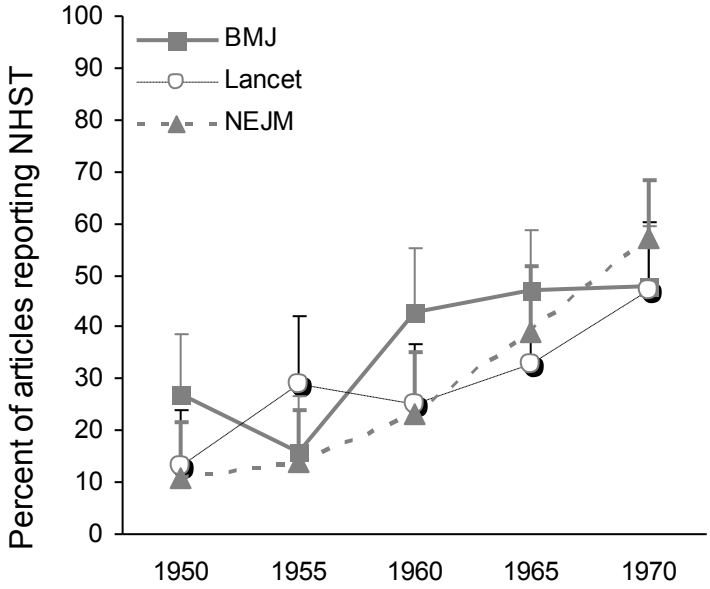


Figure 1.2. Percentage of research articles reporting NHST in *British Medical Journal*, *Lancet* and *New England Journal of Medicine* between 1950 and 1970. Error bars are upper 95% CIs (Newcombe & Altman, 2000).

By 1950 NHST was convincingly in use, if not widespread, in medicine—it was reported in 18% of articles. A relatively systematic increase in reporting saw this rate reach 51% by 1970. Obviously, results of my survey cannot directly answer the question of what happened after 1970. However, a survey by Emerson and Coditz

(1983) reported that at the end of the 1970s, 44% of *NEJM* articles reported *p* values based on *t* tests and 27% computed *p* values from contingency tables. It is not clear whether these proportions can be summed (which would indicate that 71% of articles reported *p* values by this time) or whether some articles might have instances of both procedures. In either case, it is likely the total would have exceeded the 51% of the previous decade. In medicine, however, NHST would not hold its reign for long.

In the mid-1980s there was again a shift in data analysis in medicine which saw NHST de-emphasised. (Chapter Six discusses the statistical reform of medicine in detail.) Effect size and CI reporting became routine reporting practice in many journals. In some cases (e.g., *American Journal of Public Health*) this also meant a decline in NHST reporting. In others, CIs were used in conjunction with NHST. CI reporting has remained standard practice in medicine. As I mentioned in the introduction to this thesis, recent issues of 10 leading journals[9] over 85% of empirical articles reported CIs (Coulson, Fidler & Cumming, 2005).

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

Therapeutic reform was successful and NHST was rapidly institutionalised as a routine step in clinical trial procedure. The success of the therapeutic reform was due in no small part to Sir Austin Bradford Hill. Without actually introducing the work of Fisher or Neyman and Pearson, Hill primed the discipline for the new inference methods with his *Principles of Medical Statistics*, introducing crucial concepts such as randomisation and selection bias (Hill, 1937). However, Hill himself was not impressed

---

[9] The 10 journals coded were *NEJM*, *BMJ*, *Lancet*, *Journal of the American Medical Association*, *Canadian Medical Association Journal*, *Annals of Internal Medicine*, *Archives of Internal Medicine*, *Mayo Clinic Proceedings*, *American Journal of Preventive Medicine* and *American Journal of Medicine*.

with what the new methods brought. In fact, he was amongst early critics of NHST in medicine: "the glitter of the *t* table diverts attention from the inadequacies of the fare" (Hill, 1965, p. 299).

As the quotation from Hill suggests, just over a decade after the institutionalisation of NHST in medicine, its role in clinical trials was under scrutiny. Researchers began to worry that the technique was being misused and over-relied on; that statisticians, rather than physicians, had authority over the conclusions drawn from experiments (Cutler et al, 1966). Statistical reform had begun, and by the end of the 1980s strict editorial policies (e.g., Langman, 1986; Rothman, 1986) had profoundly changed the way results were reported, if not interpreted, in medicine.

Ecology

NHST reporting in ecology increased dramatically in the early to mid 1950s: from just 6% (only 3 instances) in 1950, to 33% in 1955. From 1955, NHST reporting increased consistently to 60% in 1970. The trends in *Ecology* and *Journal of Ecology* were almost identical. Table 1.2 shows the number and percentage (with 95% CIs) of articles reporting NHST in each journal, at each time period. (Figure 1.3 also shows these percentages). The rise of NHST in ecology over these years was virtually linear in both journals.

Table 1.2.
Percentage of *Ecology* and *Journal of Ecology* articles published 1950 to 1970 reporting NHST.

	Ecology	Journal of Ecology	Total
1950	5% (2 of 39)	7% (1 of 14)	6% (3 of 53)
95% CI	1 to 17%	1 to 32%	2 to 15%
1955	34% (25 of 73)	31% (8 of 26)	33% (33 of 99)
95% CI	24 to 46%	17 to 50%	25 to 43%
1960	37% (23 of 63)	39% (16 of 41)	38% (39 of 104)
95% CI	26 to 49%	26 to 54%	29 to 47%
1965	61% (36 of 59)	52% (26 of 50)	57% (62 of 109)
95% CI	48 to 72%	39 to 65%	48 to 66%
1970	61% (71 of 116)	56% (24 of 43)	60% (95 of 159)
95% CI	52 to 70%	41 to 70%	52 to 67%

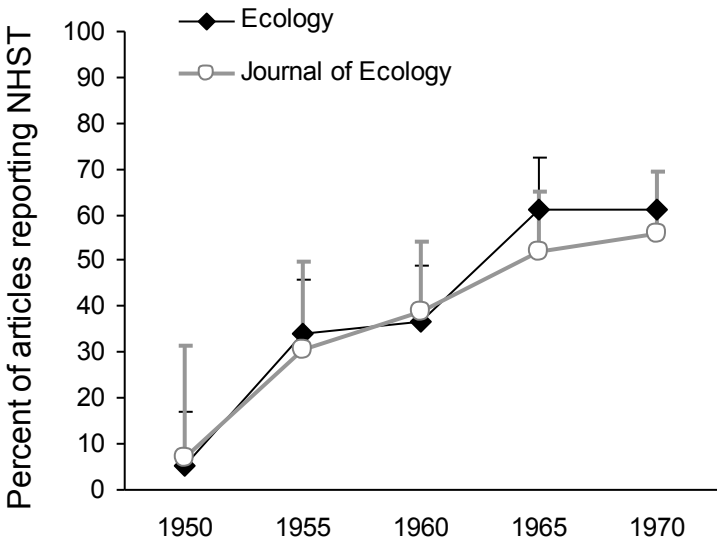


Figure 1.3. Percentage of research articles reporting NHST in *Ecology* and *Journal of Ecology* between 1950 and1970. Error bars are upper half 95% CIs (Newcombe & Altman, 2000).

Because the ecology sample was so much smaller (by necessity) than the medical sample, it was possible to make more detailed notes whilst coding. The following are some qualitative observations made during the coding of the journal *Ecology*.

In 1950 there were just two instances of NHST use. One of these was a *t* test; the other test statistic was not identified. Neither article reported statistical power or a type II error rate, but both made detailed considerations of sample size and the meaningfulness of the effect size. By 1955, however, several errors commonly associated with NHST were evident. For instance, statistically non-significant differences and correlations were interpreted as 'no difference' or 'no correlation'; in some cases they were referred to as 'insignificant.' These conclusions were drawn without any reference to statistical power, type II errors or the informal caution of 1950 articles.

Exact *p* values were virtually never reported, rather authors commonly referred to a 'significance level' or 'probability level' of 5% or 1%. Indicating statistical significance with asterisks (e.g.,* $p<.05$; ** $p<.01$) was also common reporting procedure by 1955. (Asterisks significance has been heavily criticized, e.g., Meehl, 1978; Ziliak & McCloskey 2005. This and other problems with NHST reporting are discussed further in Chapter Two.) There were also several instances of the word

'significant' being used ambiguously—that is, used in such a way that it was difficult to tell whether the author meant statistical significance or theoretical importance or practical importance. None of the articles in any year of this sample mentioned statistical power calculations.

In ecology, the use of NHST rose dramatically between 1950 and 1955 (an increase of 24 percentage points). In 1950 NHST was rarely used—only 3 instances in 53 articles—suggesting that ecology lagged slightly behind the other sciences (certainly behind psychology, and arguably behind medicine) in its uptake of the methods. By 1955, however, NHST was used in a third of articles, easily overtaking the rate of use in medicine. By 1970, 60% of ecology articles used NHST.

Sokal and Rolf (1981) reported results of a survey of *American Naturalist* articles (1890-1980) where they counted non-numerical and numerical articles. Numerical was further divided into "simple statistics employed" and "major emphasis on mathematics and statistics" (p. 5). Their results showed a linear increase in articles with a major emphasis on mathematics and statistics after 1960 (until 1980 where their survey ends). Whilst it is not possible to provide appropriate evidence, my results raise the question of whether NHST might be largely responsible for the major increase in statistical reporting found by Sokal and Rolf.

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

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

As I mentioned earlier, the uptake of NHST in ecology appears to have lagged slightly behind medicine and certainly behind psychology. The results from my survey

provide some evidence of this, but more telling is that the first successful biometry text for this audience was published in the late 1960s (Sokal & Rohlf, 1969). By contrast, popular texts in medicine introduced the new methods over a decade earlier (e.g., Mainland, 1952) and in psychology, earlier still (e.g., Lindquist, 1940).

At first it may seem odd that new methods weren't immediately transported to ecology from agriculture, where Fisher introduced his concept of ANOVA. Ecology and agriculture are related in content, at least more so than agriculture and psychology. In contrast to agriculture however, the objects that ecology studies are often not amenable to experimental control. Stigler (1999) argued that one reason for psychology's early and close relationship with NHST is that experimental psychology offered—like agriculture—the "possibility of experimental design" (p. 193). In ecology, and especially conservation biology, populations are often small and individuals may be 'cryptic' (i.e., hard to identify) and difficult to study, and conducting randomised trials is often impossible. Further, ecological studies of individuals, populations, communities and processes often take place *in situ* and require unique approaches to experimental design and control. Of course, statistical methods based on NHST were eventually adopted in ecology regardless of these difficulties and, as mentioned, continue to dominate the discipline despite their application often being controversial.

1.2.3 Psychology

NHST was fully institutionalised in psychology journals, textbooks and training programs by the mid 1950s. This history has already been documented, and preliminary theories to explain its popularity are already developed (Gigerenzer, 1989, 1993). This section is therefore a review of published data and arguments, rather than an original contribution to historical literature.

Sterling (1959) surveyed articles in four leading psychology journals[10] published in 1955 and 1956. An overwhelming 81.5% of these articles reported NHST. Almost all (97%) rejected the null hypothesis, suggesting serious publication bias at that time. Sterling acknowledged that this had probably led to a proliferation of type I errors in the psychological literature.

---

[10] *Experimental Psychology*, *Comparative and Physiological Psychology*, *Clinical Psychology, Social Psychology*

Twenty years after Sterling's survey, Rucci and Tweeney (1980) surveyed 6,457 articles published in major psychology journals. They documented the uptake of ANOVA and *t* tests between 1935 and 1952. The use of both techniques rose gradually from the mid-1930s and then declined during World War II. While the younger generation of psychologists, who had been trained in the new methods, were at war the discipline was left in the hands of an older, pre-inference revolution generation. Use of the new techniques rose again rapidly post war. By 1952, approximately 30% of articles reported ANOVA results and approximately 20% reported *t* tests.

Hubbard, Parsa and Luthy (1997) surveyed the *Journal of Applied Psychology (JAP)* from 1917 to 1994. The use of *p* values first appeared in *JAP* in 1940: A quarter (25%) of 1940-1944 articles reported them. By 1955-1959 this proportion had risen to 80.8%; in the 1990s it was 94%. Hubbard and Ryan (2000) also conducted a large survey of 12 American psychology journals; their 1950 to 1970 rates are shown with my survey results in Figure 1.1. As shown, they found a systematic increase in the use of NHST over time. For the period 1995-1998, NHST was reported in 94% of the articles they surveyed. Coulson, Fidler and Cumming's (2005) results confirm this trend continues. In articles published in 2003, in 10 leading psychology journals[11], 98% of empirical articles reported NHST.

The uptake of the NHST methods in research articles followed expository articles and new textbooks. Snedecor (1937) has been called psychology's "Fisherian prophet" (Gigerenzer, 1987, p. 19). Although officially still writing for an agricultural audience, he translated the difficult mathematics of Fisher's text into language that psychologists could understand (or, as we shall see in Chapter Two, misunderstand). Most importantly, he filled in many steps in the methodological process that Fisher had assumed. Snedecor also established the *F* statistic as the main outcome of ANOVA, rather than the critical ratio Fisher had used (Huberty & Pike, 1999).

After years of conflict between Fisher and Neyman and Pearson, psychology textbooks began anonymously presenting an incoherent hybrid of the two theories (Gigerenzer, 1987). There were few textbooks produced during World War II. One exception was Lindquist's (1940) *Statistical Analysis in Educational Research*,

---

[11] *Cognitive Psychology, Journal of Personality and Social Psychology, Journal of Consulting and Clinical Psychology, Journal of Experimental Psychology—General, Child Development, Journal of Abnormal Psychology, Cognition, Psychological Science, Journal of Abnormal Child Psychology, Journal of Health and Behavior.*

considered by some to be responsible for starting the trend of hybrid NHST presentation (Gigerenzer, 1987; Rucci & Tweeney, 1980). Lindquist referenced Fisher, and not Neyman and Pearson, despite the methods he presented being clearly Neyman-Pearsonion in origin—for example, he discussed type I *and* type II errors.

Huberty (1993) examined 28 statistical texts, written for behavioural scientists, published between 1910 and 1949. The six pre-1920 texts he reviewed focused primarily on descriptive statistics discussing, for example, mean, median, mode, quartiles and standard deviation. Of 10 texts published during the 1920s, half were also descriptive; the other half included discussions of probability and precision of estimates. Four of the five texts that went beyond descriptive statistics did so be including Karl Pearson's chi-squared goodness-of-fit tests. In the 1930s, only one textbook of eight remained purely descriptive. Fisher was referenced in two; the remaining six continued to focus primarily on K. Pearson's goodness-of-fit tests. There were four new offerings in the 1940s: one explicitly referenced Neyman and Pearson; one was consistently Fisherian; two referenced Fisher, but presented Neyman-Pearson methods (Lindquist was one of the latter).

Huberty concluded: "It took about 15 years—from 1935 to 1950, roughly—for the Neyman-Pearson philosophy to be integrated (to some extent) into presentations of statistical testing in behavioral science statistical methods textbooks." (p. 323). Also noteworthy is that by the early 1950s, half of US psychology departments were requiring courses in statistics as part of their graduate programs (Rucci & Tweeney, 1980).

The Triumph of the Aggregate

One important development that paved the way for the inference revolution in psychology was the shift from individual to group data. Danziger (1990) called this "triumph of the aggregate" (p.68).

For roughly the first four decades of the 20th century the new discipline of experimental psychology was split by two approaches: Wundtian and Galtonian. The Wundtian school's interest was 'psychic casuality' (which was a psychological version of physical causality or physiological causality). They considered the mind a synthetic unity, constituted by the processes of psychic causality, and consequently, the research unit was a single individual engaged in systematic introspection. Any further subjects were considered replications; these replications were an attempt to assess generalisation

to a population. The Wundtian school showed little interest in the inferential methods of Fisher and Neyman-Pearson. They were of little application to their studies of single subjects.

The neo-Galtonian school, which came to dominate psychology, concerned itself (as had the earlier Galtonian school) with the distribution and covariation of characteristics in populations—first with naturally occurring populations and later, with experimentally manipulated groups. The former were 'natural' in the sense of being pre-existing, for example, university students, men and women. These populations were compared on the basis of measured characteristics, such as intelligence. This school found a lucrative market in the rising industry of educational professionals and administrators: "In the United States the needs of educational administration provided the first significant external market for the products of psychological research in the years immediately preceding World War I" (Dazinger, p.103). After the war there was even higher demand—a virtual explosion of the industry devoted to intelligence and mental aptitude tests.

With such demand it is not surprising that the neo-Garltonian school grew, bringing with it an ever increasing interest in group data. Danziger reported that between 1914 and 1951 the reporting of group data in the *American Journal of Psychology* rose from 25% to 80% and the reporting of individual data dropped from 70% to 17%. In *Psychological Monographs* the trend was essentially the same. Even in journals already more Galtonian in style, such as the *Journal of Educational Psychology*, the trend was also evident: In 1914 three quarters (75%) of articles reported group data and by 1936 the proportion was 94% (Danziger, ch 5). The shift towards group data was an important first step in the inference revolution: "The triumph of the aggregate was one step in constructing a discipline ready for the introduction of formal inferential statistical techniques." (Dazinger, 1987, p. 41).

Randomisation and the Treatment Group

The concept of the treatment group helped overcome the limitations of correlational designs that continued to plague the neo-Galtonian school after the shift to aggregate data. As experimental psychology began to look to physics as a model, it became increasingly important to go beyond studies of *relations* to studies of *causation*, and to work toward uncovering psychic causal laws, equivalent to the laws of physics. In the 1930s the introduction of treatment group offered this possibility:

> To be able to make causal inferences it is necessary to introduce a comparative perspective and to study the difference in the performance of two or more groups exposed to different conditions. Thus is born a fundamentally new entity in psychological research, namely, the treatment group. (Dazinger, 1987, p. 41).

With the treatment group came the possibility of randomised trials, and the application of Fisher and Neyman and Pearson methods for testing the differences between groups on mental phenomenon quickly came to dominate the discipline.

Statistics That 'Solved' the Theory Crisis

In 1969 Morrison and Henkel wrote: "It is the social scientist's lack of theoretical development and of theoretical concern that make significance tests attractive" (p. 369). It was perhaps not only the lack of theoretical development that made NHST attractive, but the false notion that it was itself a surrogate for theory. As Danzinger pointed out, such a surrogate would have been most appealing to the young discipline of experimental psychology, which had never managed to forge strong or particularly testable theories.

> It [NHST] was a practice that reduced the demands made on psychological theorizing—no trivial achievement for a discipline that had never been able to get its theoretical house in order (p. 154).

Paul Meehl made essentially this same argument in 1967 and again in 1978; Ronald Serlin repeated in 1987. Gigerenzer (1998) explained: "Null hypothesis testing provides researchers with no incentive to specify either their own research hypotheses or competing hypotheses" (p.200). Merely identifying statistically significant differences between groups or conditions does not make a theory, particularly when those differences are only ever tested for statistical significance against zero. Yet, the illusion of theory it creates is precisely what made it so attractive to the emerging discipline.

Statistics That Served Determinism and Objectivity

Gigerenzer (1987) argued that NHST served two main ideals of the emerging 20th century experimental psychology: determinism and objectivity. As I mentioned earlier, the experimental psychology of the late 19th century modelled itself on classical physics. This placed determinism and objectivity at the centre of the new science, and these remained its tenets through the first half of the 20th century.

How was psychology able to use probabilistic thinking in service of these principles? To serve the determinism, experimental psychology rejected the use of probability at the level of theory construction; quantum physics, by contrast, embraced it. In psychology, for example, differences between individuals were treated as equivalent to errors of measurement.

> Probabilistic thinking was not tolerated as a model for how man functions; it was tolerated and used in the spirit of Laplace, as an expression of the experimenter's ignorance… [it] seldom threatened psychological determinism. (Gigerenzer, 1987, p.16).

NHST also provided experimental psychology with the illusion of a mechanized knowledge building process: In this way, it served the ideal of objectivity. The dichotomous decision procedure of NHST seemingly removed experimenter judgment from the inferential process. It appeared no longer necessary to make subjective decisions about whether a phenomenon was real, or an effect important. 'Statistical significance' became a substitute for both decisions. This rhetoric of objectivity was extremely important in psychology's struggle to be seen as a scientific discipline (John, 1993).

Before the *t* test and ANOVA, journal articles in psychology contained mostly descriptive statistics; judgments were based on eyeballing curves. Gigerenzer (1987) reported that in the *Journal of Experimental Psychology* in1925 data description and inference were often not distinguished. Inference only became a distinct and dominant practice after the mechanised NHST process was established. Gigerenzer lists three factors that enabled the spread of NHST in psychology: anonymous presentation in textbooks, neglect of alternative methods and the endorsement of journal editors. These are discussed below in turn.

In textbooks NHST was presented as a single, anonymous theory of inferential statistics. Presenting statistical theory anonymously lent credibility to the theory. It was not simply the theory of particular person, but rather "the" theory. Gigerenzer reported that 21 out of the 25 texts he surveyed did not report the names of Fisher or Neyman and Pearson. Presenting 'statistics' anonymously also allowed for its controversial history to be covered up. The method presented was a hybrid of a number of contradictory theories of probability, most notably of Fisher and Neyman-Pearson, with Bayesian overtones. Problems stemming from these contradictions went unacknowledged.

The first Bayesian texts for the social sciences did not appear until the 1970s. They made little impact on psychology as the hybrid was fully institutionalized by this time. Ignoring alternative theories was not typical of the psychological research community, which was well used to controversy—for example, the split of behaviourist and cognitive psychology. This neglect of alternatives appears to be unique to the statistical analysis.

Statistically significant results became an institutionalized criterion for publication. An often cited example of this is Melton's (1962) policy at the *Journal of Experimental Psychology*. Below Melton discusses the journal's decision to move from a publishing criterion of $p<.05$ to $p<.01$.

> In editing the *Journal* there has been a strong reluctance to accept and publish results related to the principal concern of the research when those results were significant at the .05 level, whether by one- or two-tailed test! This has not implied a slavish worship of the .01 level or any other level, as some critics may have implied. Rather, it reflects a belief that it is the responsibility of the investigator in science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounced (p.553).

The Limits of Psychology's Revolution

The impact of the new methods on experimental design was perhaps more subtle than we might expect. ANOVA did not, as one might be tempted to think, suddenly herald multi-factor experimental designs. Such designs were already in use in the practical work of intelligence testing in education and industry (Dazinger, 1987) and "in studies of reaction time, fatigue and certain cognitive skills such as reading shorthand" (Lovie, 1979, p. 153). But analysis of data from such designs had been informal and piecemeal. What ANOVA offered was a formal procedure for expressing differences between treatment and control groups.

In some research areas, the new methods made very little, if any, impact. At the very least, not *all* psychologists were immediate converts to the new methodology. Skinner (1956, 1971), for example, actively rejected them, as did other well known schools and individuals. As Danziger (1987) said:

> We shall look in vain for the psychoanalyst who suddenly becomes enamored by statistical tests of significance; we shall not find that the Gestalt psychologists embraced analysis of variance (ANOVA) as the answer to the prayers, nor shall we catch Jean Piaget employing the *t*-test (p. 36).

However, it is clear that the majority of psychology was won over by the end of the 1940s, and to a far greater extent than either medicine or ecology. Post WWI, the then products of psychological research—most notably intelligence and educational tests—were, for the first time, in high demand. NHST provided a means of mechanising the production of psychological research, even if as Danzinger points out, "it can now be seen as having contributed nothing of either practical or theoretical value" (1990, p.154). This climate of demand may have been parallel to the climate of demand in medicine in the 1950s, when pressure for increased production of pharmaceutical drugs led to the institutionalisation of clinical trials and inferential statistics in that discipline. However, unlike medicine, it is unlikely that this external pressure was the determining factor in the uptake of NHST psychology. As I have explained, for psychology, there were other aspects of NHST, beyond its efficiency, that made it so attractive—namely its ability to be (readily and problematically) substituted for theory and the illusion of objectivity generated by arbitrary criteria for statistical significance.

1.3 Conclusion

It is at least superficially curious that NHST, with its origins in agriculture and eugenics, dominated psychology so rapidly. But psychology offered something ecology, and perhaps other disciplines, lacked: "the possibility of experimental design" (Stigler, 1999, p.193). In psychology, the 'possibility' had to some degree existed since the advent of psychophysics in around 1860, but it became especially pronounced with the rise of the aggregate data and the introduction of treatment group in the 1930s (Dazinger, 1987). There at least two important explanations of why these particular inference techniques were welcomed so warmly in psychology. First, they could be readily (however wrongly) substituted for theory, 'solving' psychology's theory crisis; and second they provided the illusion of objectivity and a scientific rhetoric for the emerging experimental discipline.

In the 1930s and early 1940s, when expository articles on ANOVA were beginning to penetrate the psychological literature, medicine showed little interest in these techniques. This disinterest lasted until after World War II, when the inference methods entered medicine, largely through the pressure of pharmaceutical testing. This was a time of "unprecendented production of new synthetic drugs and of antibiotics" and "an informed public… eagerly awaited a verdict on their merits." (Hogben, 1957, p.28)

In medicine, randomised trials were not seriously adopted until the pressure of drug testing forced it—and the randomised design was a necessary prerequisite to the new inferential statistics. Once the design was institutionalized, the statistics soon followed. In ecology, randomisation proved more difficult and often still remains a serious obstacle to inference. Nevertheless, NHST eventually took the throne in this discipline too, and is yet to be deposed.

Strangely perhaps, given its history, psychology has been at the forefront of the movement *against* NHST. Criticisms began earlier in this discipline than others, and outweigh any other discipline's efforts. Yet, so institutionalised is the mechanised process in this discipline, that reformers have made only very limited impact on researchers' practices and what appears in the journals.

2
WHAT IS WRONG WITH NHST ANYWAY?

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! (Cohen, 1994, p. 997)

Before tackling the central historical issues of what reform efforts were made in each discipline and of what impact those efforts had, it seems necessary to briefly review the common criticisms of NHST. None of these criticisms are new. They have all been identified before. In fact, they have also been extensively reviewed before (Nickerson, 2000; Kline, 2004). Yet, a convincing argument that the persistence of NHST, in the face of such criticism, is a sociology of science curiosity, could hardly be made without explaining and evaluating the criticisms themselves. This chapter is divided into two distinct parts. First I provide a catalogue of common criticisms of NHST. Second I review some defences of NHST.

2.1 Criticisms of Null Hypothesis Significance Tests

Huberty and Pike (1999) divided NHST criticisms into the following categories: *Logic*, *Relevance*, *Interpretation, Use*[12] and *Alternative Analyses*. I've borrowed their category headings, though perhaps not drawn boundaries in the same places. Many of the criticisms reviewed here are as made by psychologists. This should not imply that they are irrelevant to other disciplines or have not also been made by medicos, ecologists, statisticians or others: it merely reflects the massive literature on this topic in psychology.

[12] I have relabeled this category 'misuse'.

2.1.1 Logic

… False beliefs may not be solely the fault of the users of statistical tests…This is because the logical underpinnings of contemporary NHST are not entirely consistent (Kline, 2004, p.9)

Critics of NHST logic identify problems above and beyond those that arise from researchers' misinterpretation or other shortcomings of practice. Some argue that NHST is frequently misinterpreted *precisely* because its logic is inherently flawed—that any interpretation resulting from its confused procedure will, by definition, be a misinterpretation (Falk & Greenbaum, 1995; Tyron, 2001). Others remain convinced that there are no logical flaws, only problems of *interpretation* (Abelson, 1997; Mulkai, Raju & Harshman, 1997; Reichardt & Gollob, 1997) and *relevance* (Kirk, 1996)—these are discussed later. Perhaps the best known 'logic' critics are the Bayesians.

Bayesian Critiques

Bayesians argue—and few dispute—that the probability desired by scientists is not the probability of the data, given the hypothesis: $P(D \mid H_0)$. More relevant and useful is the probability of the hypothesis, given the data: $P(H_0 \mid D)$. However, the latter requires the introduction of prior probabilities. How prior probabilities are generated is perhaps the greatest point of contention between the Bayesian and Frequentist schools. Frequentists (usually) reject the notion of priors because of the belief that priors are subjectively generated. That is, they believe priors can only be set by one's opinion or degree of belief in the likelihood of an event. Such subjectivity is not tolerated by strict Frequentists. To reject Bayesianism on these grounds, however, is to reject a straw man. Whilst some Bayesians find no difficulty in adopting subjective priors, many others do. There is a spectrum of views that fall under the label 'Bayesian'. In fact, the diversity of views led Ferson (2005) to write: "Bayesians are like snowflakes… each Bayesian is a unique assemblage of opinions about the proper methodology for statistics." (p. 6). Although some are subjective, others, for example, Harold Jeffreys (1961), promote an 'objectivist' approach to Bayesian inference. Priors in this case can be set on the basis of earlier data collection, or left uniform in the absence of such information, thereby producing an outcome similar to frequentist analysis.

There is also a fundamental difference between Bayesians and Frequentists in the understanding of 'probability'. Bayesianism is a philosophical tenet that

probabilities refer to degrees of belief or the plausibility of statements, and Bayes Theorem offers a means to updating beliefs as evidence accumulates. Bayesians combine prior knowledge, characterised by a prior distribution, with empirical data to form a posterior distribution. A posterior distribution, like a frequentist sampling distribution, is used to calculate probability values and interval estimates—but the interpretation of these probability values and interval estimates is very different to a Frequentist interpretation. Frequentists assign probabilities to events based on the long run relative frequency of their occurrence: Probability statements are only meaningful in the context of repeated sampling. Probabilities of single specific events cannot be interpreted within this framework. Further, Frequentists assign probability values to random variables (ones that can be sampled over and over in a long run series of experiments) but not to parameters (which they take to be true or fixed for the population). In contrast, Bayesians rarely distinguish between random variables and parameters.

Most Bayesians would agree with all the criticisms of NHST that follow in this chapter. They would agree, for example, that frequentist *p* values are regularly misinterpreted in the various ways listed. They would, in addition, argue that even when correctly interpreted they offer little to the progress of science, that far more useful to the progress of science is probability generated by Bayes Theorem.

For Bayesians, Frequentist CIs offer no solution to the problem. A strict Frequentist interpretation of a CI allows only statements about the long run, and not about a single interval's probability of containing the true parameter value. For example, a Frequentist can claim of a 95% CI around a mean that '95 times out a 100 an interval calculated in the way this one has been will contain the true population mean', but not that 'there is a 95% chance that *this* interval contains the true population mean'. The latter interpretation can be applied only to Bayesian credible intervals, as calculated from a posterior distribution.

Advocates of Bayesianism have been relatively marginal in psychology and medicine but have, nevertheless, had a consistent presence in the reform debate since its inception (e.g., Rozeboom, 1960). In ecology, on the other hand, Bayesians have been amongst the most outspoken critics of NHST (Ellison, 1996; Spiegelhalter et al, 2002; Wade, 2000) so too have proponents of Likelihood and information theoretic methods (Anderson et al., 2000; Burnham & Anderson, 2002). As their computational difficulties become less daunting, with faster computers and better-developed software,

these methods will no doubt increase in popularity. However, they are yet to be incorporated in most undergraduate curricula and mainstream training in any discipline, so uptake is unlikely to be rapid.

Although rejected in principle by Bayesians, some would agree that frequentist CIs have some advantages over frequentist *p* values. For example, they may agree that having information on precision is better than not having it. A short term advantage of CIs over Bayesian methods may simply be their familiarity. In many cases, CIs could be implemented immediately as a step towards remediating current problems. Widespread adoption of Bayesian methods, on the other hand, seems unlikely in the very near future, however attractive the benefits. Bayesian credible intervals may in time come to replace frequentist CIs, and may within the Bayesian framework, play the same psychological role as CIs do in the frequentist framework.

Misapplied Hypothetico-Deductive Logic

Rozeboom (1960) began as an advocate of Bayesian methods for psychology. However, his criticism of the logic of NHST—of misapplied hypothetic-deductive logic—can perhaps be viewed as beyond the Frequentist-Bayesian debate. Rozeboom (1997) described a standard model of how science proceeds under the hypothetic-deductive logic:

> We determine that uncertain proposition *C* (a possible experimental outcome or other data prospect) is logically entailed by uncertain hypothesis *H*;
>
> We ascertain that *C* is (a) true, or (b) false;
>
> We conclude that *H* has been respectively (a) confirmed, or (b) disconfirmed. (p.337)

The structure of the reasoning parallels a *modus tollen*s argument, the argument structure that provided the backbone of Karl Popper's (1968, 1969) falsificationism. The *modus tollens* in its most simple form is as follows:

> If A then B.
>
> Not B.
>
> Therefore, Not A.

The *modus tollens* argument is a valid form of reasoning when the premises are categorical (as above). However, if the first premise is a probability statement, the argument becomes invalid. For example:

If A then *probably* B.

Not B.

Therefore, *probably* not A.

This, argue 'logic' critics such as Rozeboom, is precisely the problem that afflicts NHST. To use the *modus tollens* argument probabilistically is a misapplication of the argument structure. In NHST the argument structure is:

If H_0 is true, then these data are unlikely.

These data occurred.

Therefore, H_0 is unlikely (Reject H_0).

The flaw here may not be immediately obvious—which is, of course, also part of the problem. It is clearer in some contextualised examples than others. An obvious, and perhaps the most well-known, demonstration of the flaw comes from Pollard and Richardson (1987) and was reproduced by Cohen (1994).

If a person is an American, then he is probably not a member of Congress.

This person is a member of Congress.

Therefore, he is probably not an American.

This argument again follows the *modus tollens* structure, but because the first premise is probabilistic, the conclusion is false—and obviously so!

The probabilistic *modus tollens* (despite the problem described above) is an integral part of the standard hypothetico-deductive model of science. The model permeates psychology, ecology, medicine and many other sciences, justifying itself by appeal to the authority of Popperian falsificationism. Yet as Meehl (1978), Rozeboom (1960, 1997), Cohen (1994), Oakes (1986) and others have pointed out, the resemblance between NHST, and its use of the *probabilistic modus tollens*, and Popper's philosophy of science is superficial at best. (This issue is discussed further in Chapter Seven.)

Tail Area Probabilities

Another criticism often catalogued under 'logical flaws' relates to use of tail area probabilities (Good, 1982). Using tail areas to calculate probability values requires postulating the probability of, not only the event that has occurred, but also the probability of events that have not, and potentially will not, occur. That is, we undertake to calculate the probability of not only 'this data' but also 'some more extreme data' (given the null hypothesis). Why, ask critics like Good, should our

probability value be influenced by the more extreme cases when they are not what have occurred?

Other Problems with Logic

Berkson's (1942) criticisms of NHST are sometimes put in the 'logical flaws' category by reviewers (e.g., Huberty & Pike, 1999; Nickerson, 2000). Berkson's first complaint is that the mere infrequency of an event's occurrence is not sufficient for disproving a hypothesis:

> 'If *A* is true, *B* will happen sometimes: therefore if *B* has been found to happen, *A* can be considered disproved.' There is no logical warrant for considering an event known to occur in a given hypothesis even if infrequently, as disproving the hypothesis. (1942, p. 326).

Berkson is right in that infrequency does not equal impossibility. Yet, the criticism he makes here would perhaps apply to any kind of probabilistic reasoning. In short, he sets the bar too high. Berkson was also concerned that the logic of NHST did not agree with the reasoning of ordinary discourse, or with reasoning scientists used in the scientific laboratory. The main difference he identified between NHST logic and the procedural logic of laboratories was that the former was engaged only in disproving things. Of NHST logic he wrote:

> But ordinarily evidence does not take this form. With the *corpus delicti* in front of you, you do not say, 'Here is the evidence against the hypothesis that no one is dead.' You say, 'Evidently someone has been murdered.' (p.326).

Again, Berkson's criticism seems to range beyond what *any* statistical procedure can reasonably held accountable for. In fact, his criticism here could apply to the entire philosophy of falsificationism. Indeed, these charges of asymmetry may be more appropriately directed at Popper than at any particular school of statistical inference.

Other issues identified in 'logical flaws' sections of reviews of NHST critiques seem to me not so much inaccurate, as miscatalogued. For example, it is difficult to see why the fact that researchers being "given little guidance in setting levels for alpha and beta" (Huberty & Pike, 1999) earns a place under the heading 'logic'. Admittedly, with some criticisms it is difficult, and perhaps not helpful, to catalogue them as either 'logical flaw' or 'misinterpretation'. The two can be inextricable, as in the case of the inverse probability fallacy (classified here as 'misinterpretation') and the argument of

misapplied hypothetico-deductive logic ('logic'). Where criticisms of logic meet those of relevance is also somewhat fuzzy, as we shall see.

2.1.2 Relevance

Cohen claimed that even when NHST is used and interpreted correctly "it is not the way any science is done" (1994, p.999). In fact, it is the way much science is done, but perhaps not the way it ought to be done. Rozeboom (1960) also believed NHST was "seldom if ever appropriate to the aims of scientific research" (p.417). Carver (1978) too expressed this sentiment: "Even if properly used in the scientific method…research would still be better off without statistical significance testing." (p.398).

What would be more relevant to the goals of science? Along with many other critics of NHST in various disciplines, I believe that estimation of effect sizes should be the primary outcome of research, and that science would have progressed further had this been the case. As Kirk (1996) argued:

> How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between *A* and *B* and the error associated with our estimate; knowing *A* is greater than *B* is not enough (p. 754).

Kirk's comments reflected those made earlier by Tukey (1991): "the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'Are the effects different?' is foolish" (p.100). Similarly, Demming (1974) said: "The question in practice is never whether a difference between two treatments A and B is significant." (Invited talk at Princeton Unviersity on November 17th 1974, cited in Salisburg, 2001, p.255).

Even when not promoting estimation of effect size, many critics emphasize that the probability value (*p* value) that results from NHST does not tell scientists the probability they want to know (Cohen, 1994; see the epigraph at the start of this chapter). As I have explained, the NHST *p* value provides the probability of getting this particular experimental result or one more extreme (i.e. further into the tails of a null distribution), given that the null hypothesis is true: $P(D \mid H_0)$. However, the question of genuine interest is usually not "how likely our data is *given* some hypothesis?" but rather, "how likely it is that our hypothesis is true given some data?" The latter is

$P(H_0 \mid D)$, the solution to which is provided only by Bayes Theorem. This argument leads *both* Bayesian critics and others to conclude that the probability value provided NHST, interpreted literally, is a relatively useless one for the purposes of theory building or cumulative knowledge growth.

2.1.3 Interpretation (A Catalogue of Misconceptions)

Misconceptions associated with NHST are well documented in surveys of statistical reporting in journals (Dar, Serlin & Omer, 1994; Finch, Cumming & Thomason, 2001; Kieffer, Reese & Thompson, 2001) and in direct studies of researchers' understanding (Haller & Krauss, 2002; Oakes, 1986; Tversky & Kahneman, 1971). The misconceptions listed below are specific false beliefs about the meaning of a *p* value and/or the concept of statistical significance. Before listing these false beliefs, it may be useful to review some definitions.

Kline (2004) offered three statements that may be considered correct interpretations of $p<.05$:

> 1. The odds are less than 1 to 19 of getting a result from a random sample even more extreme than the observed one when H_0 is true.
> 2. Less than 5% of test statistics are further away from the mean of the sampling distribution under H_0 than the one for the observed result.
> 3. Assuming H_0 is true and the study is repeated many times, less than 5% of these results will be even more inconsistent with H_0 than the observed result. (p.63)

All three perhaps need to be supplemented with "given no measurement error, and only random noise" to be technically correct. However, even with their own imperfections, they can be contrasted starkly with the incorrect definitions, misconceptions, that follow.

ia). The p value is the Probability that the Null Hypothesis is True, Given the Data
ib). The p value is the Probability that the Null is True

These might be two misconceptions, though the second is probably simply an abbreviated version of the first. 'P is the Probability that the Null Hypothesis is True, Given the Data' is a simple confusion of two conditional probabilities, $P(D|H_0)$ and $P(H_0|D)$. Taken literally, statement 1a) 'P is the Probability that the Null is True', fails

to recognise that the probability given by NHST is a conditional probability. $P(D|H_0)$ is erroneously converted to $P(H_0)$, probably via $P(H_0|D)$.

In fact *p* values provide no *direct* information about the truth or falsity of the null hypothesis, conditional or otherwise. This misconception is thought to stem from what has variously been called: 'the inverse probability fallacy' (Tversky & Kahneman, 1971, and subsequently used in this thesis); 'the permanent illusion' (Cohen, 1994); 'confusion of the inverse' (Dawes, 1988); 'the fallacy of the transposed conditional' (Diaconis & Freedman, 1981); and 'the illusion of probabilistic proof by contradiction' (Falk & Greenbaum, 1995). This is perhaps in turn related to the 'premise conversion error' of conditional logic, which equates 'If P then Q' with 'If Q then P' (Dawes, 1988).

How common is the *p* value misconception rising from the inverse probability fallacy? Oakes (1986) reported that over one third (36%) of 70 academic psychologists agreed with a direct statement of inverse probability: "The probability of the null hypothesis has been found" (p.80). Not only did many researchers agree with this statement as a plausible interpretation of *p*, but almost half (46%) described it as typical of their usual interpretation of NHST results. When statements of the fallacy were slightly less explicit, even more researchers slipped. For example, 66% agreed that "The probability of the experimental hypothesis can be deduced" and 86% agreed that "The probability that the decision taken is wrong is known" (p. 80).

Haller and Krauss (2002), repeating Oakes' survey in 6 German universities, found that almost 20 years on, the misconception was diminished by only approximately 10 percentage points. In 2002, 26% (10 of 39) of academic psychologists agreed with the direct statement: "The probability of the null hypothesis has been found." Perhaps not surprisingly, more students (32%, 14 of 44) demonstrated the misconception. What is surprising about Haller and Krauss' study, however, was the 17% (5 of 30) of the methodology instructors they surveyed also demonstrated the misconception. For less explicit versions of the fallacy, the frequency of misinterpretation again rose. For example, a third (33%, 13 of 39) of academic psychologists and a third of methodology instructors (33%, 10 of 30) agreed that "You can deduce the probability of the experimental hypothesis being true" (p. 5); 59% (26 of 44) of students agreed too.

Available evidence suggests that this particular misconception may be difficult to extinguish. Falk and Greenbaum (1995) studied the effect of directly confronting this

misconception. Students (n=55) were given a short version of Oakes' survey after being assigned Bakan (1966). Bakan (1966) directly warns against interpreting p=.05 as meaning there is a 5% chance of the null being correct and a 95% chance of the alternative hypothesis being correct. After being directly instructed on the fallacy and reading Bakan's article, 79% of Falk and Greenbaum's students still committed the fallacy, agreeing with a statement that a statistically significant result demonstrates that H_0 is improbable.

How serious, for science, are the consequences of committing this fallacy? Nickerson (2000) outlined two conditions under which $P(D \mid H_0)$ and $P(H_0 \mid D)$ produce roughly the same answer. If these conditions are typical of most applications of NHST, then we can expect that this misconception has not had a serious negative impact on conclusions drawn from such analyses. The first condition under which the two probabilities will be equal is when the prior probability of the null and the prior probability alternative are at least roughly equal (or the alternative more likely). In typical practice in psychology, ecology and medicine, the null hypotheses are formulated as *nil* nulls, meaning hypotheses of no difference, no impact, no correlation etc. Almost all nulls of this nil kind are *a priori* false, which means that the prior probability of the alternative hypothesis would often be higher than that of the null. In these disciplines then, Nickerson's first condition seems satisfied.

The second condition laid out by Nickerson is that the conditional probability of the data, given the alternative hypothesis, $P(D \mid H_a)$, be much larger than the conditional probability of the data, given the null, $P(D \mid H_0)$. Again, the assumption that this is true is itself often the motivation for empirical investigation, and data collection, in the first place.

Given that the two conditions seem typical of most research in at least psychology, medicine and ecology, perhaps the consequences of this misconception are after all not so serious? If Nickerson is right that under these conditions, it is reasonable to let $P(D \mid H_0)$ be a proxy for $P(H_0 \mid D)$ then perhaps conclusions based on the inverse fallacy are not so wildly off the mark, and science has not been led so far astray. It is difficult to know for sure.

However, there is another, less direct, consequence of this misconception to be considered. Oakes (1986) used widespread endorsement of the inverse fallacy to explain the neglect of statistical power in the psychology research literature. His

argument can be paraphrased: Why would researchers bother about calculating the probability of detecting an effect of a given size (statistical power), when they (believe they) already know the probability of the null hypothesis being true? (see Oakes, p.82-83). Accepting Oakes' argument makes the misconception a serious issue. Statistical power is rarely considered or reported in research articles (Anderson et al, 2000; Finch, Cumming & Thomason, 2001), and the consequences of illusory inconsistencies in research literature are serious, as is explained later in this chapter.

ii). 1-p is the Probability of the Alternative Hypothesis being True

This misconception is closely related to 1; n fact, it is entailed by 1. A *p* value does not provide direct information about the truth or falsity of the alternative hypothesis for reasons already given—although many students, researchers and even methodology instructors believe it does (Haller & Krauss, 2002; Oakes, 1986).

iii). The p value is the Probability that the Results are Due to Chance

Carver (1978) called this the 'odds-against-chance fantasy'. Kline (2004) explained it as the false belief that p=.05 means there is 5% likelihood that the results are the product of chance alone. It also manifests in interpretations of the *p* value as a measure of sampling error (Finch et al, 2001).

In fact, the 5% type I error rate has a much more specific interpretation. It is the probability of falsely rejecting the null, when the null is in fact it is true. The odds against chance fantasy may lead to ignoring the usually much higher chance of making a type II error because one believes 5% is the overall error rate. Furthermore, this misconception could contribute to the common failure to control for spurious statistically significant results in studies where multiple hypotheses are tested. In a study testing 20 null hypotheses (which is not uncommon in many disciplines) we should expect at least one test to lead to rejection of the null, even though the null is true. The problem of inflated type I error rates is well known, and many critics have called for increased use of Bonferroni adjustments and related procedures (see Bland & Altman, 1995).

iv). The p value is an Inverse Indicator of Effect Size

Because *p* values are a function of both sample size and effect size, neither can be read directly from a *p* value. Kline (2004) called the belief that *p* values provide direct information about effect size the "magnitude fallacy" (p.66). Nickerson (2000) provided an example: "I recently reviewed a manuscript that described a response time that was about 16% slower than another as being 'marginally slower' than the latter, because $.05<p<.06$." (p.257). Effects of practical importance that fail to reach statistical significance are often dismissed. Similarly, very small or unimportant effects that do reach $p<.05$ are often taken to be meaningful on that basis alone.

In some articles, authors fail to provide either raw or units-free effect size measures (see Chapter Four for psychology, especially the survey of the *Journal of Consulting and Clinical Psychology*; see Chapter Eight for ecology, especially the survey of *Conservation Biology* and *Biological Conservation*). This reporting deficiency is plausibly a result of the magnitude fallacy: Why bother reporting an effect size when you believe the *p* value to be one?

v). The p value is an Inverse Indicator of the Probability of Replication

Sometimes called the 'replicability fallacy', this is the false belief that a *p* value of .05 means that 95 times out of 100, the observed statistically significant difference will hold up in future investigations. In Oakes' survey, 60% of researchers agreed with this statement of the replicability fallacy:

> You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. (p.79)

In Haller and Krauss' update of Oakes' study, 37% of methodology instructors, 49% of academic psychologists and 41% of psychology students agreed with the statement. If there is any true decrease in the hold of this fallacy since Oakes it is disappointingly slight, and the misconception's pervasiveness amongst methodology teachers is alarming.

Carver (1978) claimed that replication largely remains an empirical question, to be answered only by future studies. That was perhaps the case in 1978. However, computer intensive resampling strategies have made replication a question that can also be addressed by simulation, for example, bootstrapping, jack-knifing and other MCMC (Monte Carlo Markov Chain) methods. Some very recent developments in measures of

replicability include Killeen's (2005) p_{rep}, which offers a statistical testing-based approach to measuring replication, and Cumming (2005), who discusses an equivalent estimation-based approach.

vi). Statistical Non-significance Means 'No Effect'

Researchers frequently report statistically non-significant results, and interpret them as evidence of no effect or no impact, usually without any reference to statistical power (Anderson et al., 2000; Parris & McCarthy, 2001). Without statistical power and effect size reports, statistically non-significant results are virtually uninterpretable. The misconception that statistical non-significance is direct evidence of 'no effect' has the potential to cause damage of various kinds in all disciplines, as we shall see Chapter Three. Conclusions based on this misinterpretation may be particularly damaging in ecology and conservation biology, where threatened or endangered populations leave little margin for recovery from this mistake (Taylor & Gerroditte, 1993).

vii). Statistically Significant Results are Necessarily Theoretically Important

A statistically significant result is not necessarily theoretically important. An effect of even trivial size will be statistically significant in a high-powered sample. Similarly, important effects can fail to reach statistical significance in poorly-designed, low-powered experiments. For this reason there have been various calls for authors to report not merely statistical significance, but also the clinical, practical or biological importance of effects (Kendall, 1997, 1999; Kirk, 1996; Anderson et al., 2000). Kline (2004) referred to the confusion of statistical and practical significance as the "meaningfulness fallacy" (p.66). This fallacy also entails the false belief that rejecting the null proves not only the statistical alternative hypothesis, but also that it proves the research and theory behind statistical hypothesis (Meehl, 1978).

2.1.4 Misuse

… where some see significance testing as inherently at fault, I believe the problem is better characterized as the misuse of significance testing (Rossi, 1997, p.175)

The over-use of, or over-reliance on, NHST is also often cited as a serious problem, particularly the simple dichotomous accept-reject decisions the procedure

often leads to. For this category of criticism, however, the neglect of type II errors and statistical power must be acknowledged as the most widespread offence.

Low and Unknown Statistical Power

In 1962 Jacob Cohen published the first survey of statistical power in the psychological literature. He calculated the average power for (arbitrarily designated) small, medium and large effect sizes of 70 articles published in 1960 issues of the *Journal of Abnormal and Social Psychology*. For medium effect sizes, thought to be typical of psychological effect sizes in that sub-discipline, the average power was .48. The situation changed little in subsequent decades. Rossi (1990) surveyed articles published in 1982 in the *Journal of Abnormal Psychology* (the same journal Cohen surveyed: The word 'social' had since been dropped from the title), the *Journal of Consulting and Clinical Psychology* and the *Journal of Personality and Social Psychology*. The average power to detect small, medium and large effects in the articles surveyed (n=221) was .17, .57 and .83. Roughly 20 years after Cohen, the average statistical power, for medium effect sizes, had risen from .48 (Cohen, 1962) to .57 (Rossi, 1990)—hardly the increase desired! Sedelmeir and Gigerenzer (1989) also surveyed the *Journal of Abnormal Psychology.* In their sample, the average power for medium effect sizes was .5—virtually identical to the average power Cohen reported for 1960. Little wonder then that Hunter (1997) compared the average psychology experiment to flipping a coin!

Unfortunately, the problem is not limited to the *Journal of Abnormal Psychology*. Rossi (1990) provided a summary table of 25 different studies, covering 40 journals, of statistical power in published articles in diverse areas of psychology and other disciplines. In only 2 journals did the average power for detecting a medium effect reach .8 or higher. Other surveys also report similar average statistical power to Cohen's original estimate (Bezeau & Graves, 2001; Clark-Carter, 1997; Kosciulek & Szymanski, 1993; Mone, Mueller & Mauland, 1996) in a variety of journals and fields including neuroscience and counselling. An exception to this general trend is a power analysis survey of health related journals (Maddock & Rossi, 2001). This survey of three journals found adequate power to detect both large and medium effects. Maxwell (2004) offers the following economic explanation of this exception:

> Research in the health-related journals tends to be federally funded, and federal funding agencies may be likely to require evidence of sufficient statistical power before deciding to fund a proposal (p. 148).

Unfortunately, *low* statistical power is not the only problem. In fact, it is not even the main problem, since meta-analysis offers the opportunity for low powered studies to make an important contribution to the research literature. Often there is no way to avoid a low power experiment. For example, when working with natural groups, one cannot infect extra patients with a rare disease, or increase the population of an endangered species. Conducting a large scale study may simply be too costly or time consuming. All of these things are understandable and justifiable. Serious problems arise only when statistical power is both low *and* unknown.

A number of journal surveys demonstrate that statistical power *reporting* rates are alarmingly low. For example, in the journals *Conservation Biology* and *Biological Conservation* 3% of 2001-2002 articles and 8% of 2005 articles reported statistical power (these percentages are calculated out of the total number of articles using NHST and reporting at least one statistically non-significant result; see Chapter Eight). In psychology, the reporting rates of statistical power are similarly low (Finch, Cumming & Thomason, 2001; Fidler, Cumming, Thomason et al., 2005).

Low and unreported statistical power has made research literature in many disciplines difficult to interpret. For example, given that in psychology the average power (for effect sizes considered typical of the discipline) is roughly 50%, it should not be surprising that one study would find a statistically significant result and the following study not. However, this often leads research programs astray with the search for illusory moderating factors (Hunter & Schmidt, 2004; Schmidt, 1996). The inability to draw conclusions from inconsistent results can lead to researchers giving up on important theories: As Meehl (1978) famously said, many theories in psychology "suffer the fate that General MacArthur ascribed to old generals—They never die, they just slowly fade away." (p.807).

In ecology, initial reform efforts placed a great focus on statistical power (Fairweather, 1991; Green, 1989; Hayes & Steidl, 1997; Peterman, 1990; Mapstone, 1995; Taylor & Gerrodette, 1993; Toft & Shea, 1983). In this discipline, power has received a lot more attention than, for example, CIs. Confusion was created, in the mid 1990s, by recommendations to use post hoc or retrospective power analysis, based on the observed effect size, published in prestigious journals (e.g., Reed & Blaustein, 1995;

Thomas & Juanes, 1996). Although this misguided practice was explicitly advocated and debated in ecology, it is not unique to that discipline. For example, SPSS[13] (commonly used in psychology and other social sciences) routinely provides retrospective power calculations based on the obtained effect size. Other programs that have (at least historically) produced retrospective power calculations based on the obtained effect size as a default include SAS, JMP and Sigmastat (Thomas, 1997).

Retrospective power using the obtained (or observed) effect size provides very little information. In fact it merely tells us what we would already know: that detecting a significant result was less likely than not. If an experimental effect is tested for statistical significance, and fails to reach $p<.05$ in a two tailed test, the retrospective statistical power of the study, for the obtained effect size, can not be higher than .5 (for a one tailed test, on the other hand, statistical power could be as high as .51!). As Thomas (1997) plainly stated, "calculating power using the observed effect size and variance is simply a way of re-stating the statistical significance test." (p.279). Retrospective power based on the observed effect also fails to contribute any new information in the case of statistically significant results. For example, it cannot help determine at what reduced sample size a result would fail (on average) to meet the criteria for statistical significance. (Retrospective power based on a predetermined important effect size, however, may be useful for this purpose, even when results are statistically significant.)

Retrospective power based on the observed effect size has now been severely and justifiably criticized (Hoenig & Heisey, 2001). However, of the few researchers who report statistical power in either psychology or ecology, even fewer indicate the origins of the effect size used in the calculation, so it is difficult to tell how widespread the practice remains. Given the default of most software programs the practice may still be common.

Ubiquitousness

NHST has dominated—to the point of exclusion—other statistical methods (Carver, 1993; Dar, Serlin & Omer, 1994; Kirk, 1996). Even amongst those who would argue that *p* values contain some worthwhile information, virtually none would agree that they tell the complete story. Yet, in many cases they are reported and interpreted as

---

[13] SPSS Inc. (2004). SPSS for Windows [Computer software] (Version Release 13.0). Chicago: Author

though they do. Discussions of results often slip into an analysis of the magnitude and importance of the effect, the probability of its replication and the probability of the hypothesis being true—all based on nothing but the *p* value.

Dichotomous Decision-making

Surely God loves the .06 nearly as much as the .05. (Rosnow & Rosenthal, 1989, p. 1277)

Meehl (1978), amongst others, argued that dichotomous decisions, based on a somewhat arbitrary cut off, are not the way science ought to be done. The real questions, he argued, are those of estimation. Further, Meehl argued, single studies should never satisfy our research questions to the extent we should be prepared to accept or reject a hypothesis after just one experiment.

However, it is often the case that decisions do need to be made and dichotomous ones at that: to close (or not) a fishery; red list (or not) an endangered species or mix (or not) the genetic populations to conserve a declining population. Often in applied ecology and conservation biology single studies will be the only basis for such decisions, either because they require immediate action or because cumulating results across studies makes no sense for experiments which have been designed specifically to assess local areas for local actions. Results may directly inform policy and action—for example, to cease or continue old growth logging, or where to locate a waste facility. Such scenarios form the more compelling defences of NHST. Yet, there remains a strong case for using CIs in these decision making contexts. CIs can be used to judge statistical significance, so no information is lost. Yet, precision and effect size information *is* gained. (Chapter Ten provides recommendations for interpreting CIs.)

Implausible Nil Nulls

NHST is often upheld as the operationalism of Popperian philosophy of science. However, testing nil nulls violates the principle of falsificationism. There is nothing 'bold' about a nil null conjecture (Meehl, 1978). In Ecology, nil nulls are equally implausible (Anderson et al., 2000; Johnson, 1999): for example, 'logging causes no decline in possum populations' or 'increased harvesting does not affect fish stocks'. Of course such actions have *some* effect—the relevant question is not "do they effect…?" but rather "how much do they effect…?" and/or "does it matter?"

Publication Bias

As early as 1951, Yates, a contemporary of Fisher's, wrote that "scientific research workers often … regarded the execution of a test of significance on an experiment as the ultimate objective" (p.33). Whilst unfortunate, it is perhaps not surprising that this was, and largely still is, the case when, as Eysenck explained less than a decade later, the word 'significant' had become "has become a shibboleth which divides the successful from the unsuccessful research" (1960, p.269).

Publication bias towards statistically significant results has been well documented for some time (Bozarth & Roberts, 1972; Sterling, 1959). The outcome of this bias has become known as the 'file draw problem', named in reference to the final destination of papers with statistically non-significant results. Such bias has the potential to dramatically distort the research literature, and provide serious problems for meta-analysts. The existence of the bias has been well documented in clinical trials (Alison, Faith & Gorman, 1996; Berlin, Begg & Louis, 1998), and in other areas, to varying extents, including biology and psychology (Bauchau, 1997; Earleywine, 1993).

Rosenthal (1979) developed the 'file drawer statistic' as an estimate of how many unreported studies with *z* scores of zero would be required to exactly cancel out the statistical significance of a given meta-analysis. Adaptations of the file draw statistic used estimates of the effect size in unpublished studies to overcome the arbitrariness of assigning a zero *z* score (Iyengar & Greenhouse, 1988; see Hunt (Ch 6), 1997). Yet, the uptake of these kinds of statistics has been far from widespread.

Fortunately, recent journal surveys suggest the file drawer problem may no longer be as serious as it once was when Sterling (1959) found 97% of published psychology articles using NHST included tests which rejected the null hypothesis (i.e., reported statistically significant results only). More recently, in psychology and ecology, roughly 80% of articles report at least one statistically *non*-significant result (see surveys in Chapters Four and Eight). However, it is still difficult to estimate how many papers do not get submitted when the *major* findings are statistically non-significant.

2.1.5 Alternative Analysis

The collective criticisms reviewed here present a strong case against NHST, raising challenges on many levels. Even if one is not convinced by arguments that the

logic of NHST is fatally flawed, it is hard to argue against claims that it is largely irrelevant to the enterprise of scientific research. It is demonstrably widely misinterpreted and misused—and half a century of criticisms have done little to change these practices in disciplines where it remains the dominant technique. This alone is reason enough for some critics to call for its abandonment (Hunter, 1997; Schmidt, 1996). The availability of alternative analyses—such as effect sizes and CIs—only serves to make the case for statistical reform stronger. I briefly listed these alternatives to NHST in the introduction of this thesis. Here I describe them again, in more detail.

Effect Estimation and Confidence Intervals

In psychology most reformers have advocated increased use of effect sizes and CIs, as either a supplement or a replacement for NHST. Harlow (1997) identified CIs as the most commonly recommended alternative to NHST by contributors to *What If There Were No Significance Tests?* In medicine, CIs were also the most common recommendation and as I mentioned in the introduction to this thesis, statistical reporting in journals now largely reflects this consensus, with around 85% of 2003 articles in 10 leading medical journals reporting CIs (Coulson, Fidler & Cumming, 2005). In the ecology reform literature CIs have received some attention (DiStefano, 2003; Wildlife Society News, 1995; Cherry, 1998), but they have not been the main focus of reform.

So, what exactly are the purported advantages of CIs? First, they make uncertainty explicit. By this I mean that CIs offer immediate information about *precision*. A wide interval indicates a lack of precision; a narrower interval, relatively better precision. CIs contain a set of plausible values for the population effect, so wider intervals rule out few values as plausible. In other words, they give a less focussed estimate of the effect. This means that studies with poor precision cannot be mistaken as evidence for nil effects, one of the major problems associated with *p* values.

Second, CIs by definition contain point estimates of effect size. A CI around a mean difference will include a best estimate of mean difference, which is the raw effect size of the study. This means that when CIs are used to report results, effect size cannot be overlooked. By contrast, when reporting *p* values it is common for researchers to neglect effect size reporting. As I have mentioned, in psychology and ecology effect sizes (including the relevant mean differences) are often missing in articles reporting *p* values.

Third, CIs do not preclude decisions. CIs can be also be used to reject or fail to reject the null, when appropriate, by noting whether or not the null is captured. Though this is far from the most useful aspect of the CIs it is important to recognise that they are capable of fulfilling statistical decision making needs when those needs exists.

In addition, the claims below have been made in the favour of CIs. Whilst plausible, some of these are yet to be empirically tested. In the final chapters of this thesis, Chapters Nine and Ten, I report results of a preliminary research program in this area.

CIs Facilitate Meta-analytic Thinking. CIs may help facilitate meta-analytic thinking (Cumming & Finch, 2001). That is, they may assist thinking across the results of independent studies, acknowledging prior information with an emphasis on effect size, rather than making dichotomous 'reject' or 'fail-to-reject' decisions based on the outcome of single experiments. In an estimation approach, where the primary research outcome is an effect size, further research improves the precision of the effect size estimate, by narrowing the interval. Of course, decisions can still be made, but the illusion of objectivity provided by NHST is removed and the uncertainty involved in the decision is made explicit. When authors fail to report effect sizes, and estimates of their uncertainty, it is often not possible for their studies to be included in quality reviews or meta-analyses.

Results presented as merely 'statistically significant' or 'not' can create the illusion of inconsistency in the literature, particularly when review studies simply tally 'significant' against 'non-significant'. The literature on the effects of toe clipping on frog survival is an example of such an apparent inconsistency. Toe clipping—removing a combination of whole or part toes—is commonly used to mark *anurans* (frogs and toads), so they can be identified when recaptured. A number of studies investigating this method have reported adverse impacts, including a reduction in the return rate of marked animals; other studies have failed to find such impacts. Parris and McCarthy (2001) and McCarthy and Parris (2004) re-analysed data from four independent, and seemingly inconsistent, studies on the effect of toe clipping on frog survival. When presented using CIs (Parris & McCarthy, 2001) and Bayesian credible intervals (McCarthy and Parris, 2004) the illusion of inconsistency in the data disappears. The reduction in return rate is consistent across studies and worrying large. Differences in 'significance' are exposed as the simple product of the relative precision of the original studies. This example is discussed again in further detail in Chapter Three. My

argument here is that CIs can help us identify patterns in data, and across studies, that statistical significance tests disguise. When presented visually, CIs may be even more effective in this role.

Determining Adequate Sample Sizes. In the many discussions and interviews I have had with researchers about whether CIs can replace NHST, I often find most agree with the arguments presented so far. Eventually, however, they are bound to ask: "But what about power? You still need power to do sample size calculations." I then explain that CIs can also be used to calculate sample size, and that in fact, this too has some advantages—the primary one being that statistical power is inextricably linked with dichotomous decision making and CIs are not.

The central question of a power analysis is about rejecting or not rejecting a null hypothesis. A power approach to sample size calculations requires the researcher to specify an effect size of interest; values for the type I error rate (e.g., .05) and for power (e.g., .80); and have available an estimate of population variance, perhaps from previous research or a pilot test. In an estimation approach, reliance on dichotomous decision making is avoided, and so is the need to specify a particular effect size—only the desired width of the interval need be specified. For the chosen experimental design, an estimate of population variance can be used to calculate the sample size needed to achieve some expected width or precision.

A power approach allows the researcher to say "If there is a true effect of 5 units, I have a .80 chance of rejecting the null hypothesis at the .05 level", whereas the precision approach justifies statements like "I expect the 95% CI to have a total width of 6 units, in other words I expect to estimate the true effect, whatever that is, to ±3 units". From a precision perspective the researcher is considering, for example, whether to use n=25 and estimate the effect with expected precision 3 units, or n=100 to achieve expected precision of 1.5 units.

Di Stefano, Fidler and Cumming (2005) explained how sample size can be calculated from CIs, and showed how to calculate an upper bound on precision, meaning that a researcher can make an additional statement like "I'm 90% sure that my experiment will give precision of 4 units or better". Kelley, Maxwell and Rausch (2003) present a CI approach to calculating sample size which they refer to as "accuracy in parameter estimation" or an *AIPE* approach. Daly (2000) discusses differences, and reasons for those differences, in sample size calculations based on NHST and equivalent

calculations based on CIs. Beal (1989), Grieve (1991) and Goodman and Berlin (1994) also deal with how to use CI to determine adequate sample size.

Cognitive Advantages. Since CIs rely on the same sample information as significance tests, and belong to the same, frequentist, philosophy of statistics, some may be tempted to think they are 'the same thing' as significance tests. Yet, they are different in important ways. The belief that they are the same ignores the extra information about precision that a CI provides, and also dismisses a mass of evidence that different formats of equivalent information can profoundly affect our ability to complete conceptual algorithms and reason using the information (e.g., Gigerenzer & Hoffrage, 1995). Richard Feynman noted this phenomenon: Some derived formulations of physical laws lead to the discovery of new laws, some do not. Even when mathematically equivalent, "Psychologically they are different because they are completely unequivalent when you are trying to guess new laws" (1967, p.53).

Theory in cognitive psychology suggests that successful formats—those that allow us to best use quantitative information to reason—have shorter *information menus* or fewer pieces of separated information (Gigerenzer & Hoffrage, 1995). CIs have this in their favour. Best practice use of NHST requires the separate reporting of (at least) effect size, *p* value and a statistical power calculation. In this format the information is fragmented, and therefore more difficult to integrate; this is not the case with CIs.

Standardised Effect Sizes

In psychology, standardised effect sizes are also a commonly recommended supplement to NHST (Harlow, 1997). By converting effects of all measurement units to units of standard deviation, they facilitate meta-analysis. CIs around standardised effect sizes are also becoming a popular recommendation (Cumming & Finch, 2001; Fidler & Thompson, 2001; Smithson, 2001; Steiger & Fouladi, 1997; Thompson, 2002), although they are as yet rarely reported in journal articles.

However, in medicine, standardised effect sizes have been severely criticised: "A key problem with such measures is that the so-called 'standard unit' used to construct them actually varies across studies, rendering them noncomparable and useless for meta-analysis." (Greenland, 1998, p. 671). Some other criticisms of standardised effect sizes are discussed in Chapter Five and this one (and proposed solutions to it) is discussed in further detail in Chapter Seven.

Likelihood and Information Theoretic Methods

One increasingly common recommendation in ecology is the use of information theoretic approaches—a move led by Burnham, Anderson and others (e.g., Anderson et al, 2000; Burnham & Anderson, 2001; Spiegelhalter et al, 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 values are used to compare competing models, and to combine, or average, models to make multi-model inferences.

In additional to expository articles, there have been several applications of these techniques in the literature (e.g., Frair, Nielsen, Merrill et al., 2004; Gibson, Wilson, Cahill & Hill, 2004; Johnson, Seip & Boyce, 2004; Tyre, Tenhumberg, Field, Niejalke, Parris & Possingham, 2003). AIC has its proponents in psychology too (Wagenmakers & Farrell, 2004) and there are rare cases of it being applied in psychological research (Boivin, Pérusse, Dionne et al, 2005; Chorpita, 2002; Muris, Schmidt, Merckelbach & Schouten, 2001). There are also some cases of AIC being used in medicine (Hutton, Cooke & Pharoah, 1994; Wong, Ko, Hui et al, 2004; Schneeweiss, Maclure, Carleton, Glynn & Avorn, 2004), though it also remains far from mainstream in this discipline.

Bayesian Methods

Again, it is in ecology that Bayesian methods received considerable attention (Clark & Lavine, 2001; Ellison, 1996; Harwood, 2000; Wade, 2000). Hillborn and Mangel's (1997) Bayesian manifesto *The Ecological Detective* made a large impression on the field (it also includes chapters on likelihood approaches). A new text by Mick McCarthy (in press) provides a practical guide to applying Bayesian methods in ecology.

In medicine Bayesian inference also has some (though relatively fewer) strong advocates (Freedman, 1996; Kadane, 1995; Spiegelhalter, Myles, Jones, & Abrams, 1999). Perhaps interest has grown in recent years. In 1998 Bland and Altman wrote:

> The Bayesians are much fewer [than the frequentists] and until recently could only snipe at the frequentists from the high ground of university departments of mathematical statistics. Now the increasing power of computers is bringing Bayesian methods to the fore (p.1151).

Medical statisticians were also responsible for developing software for Bayesian analysis. WinBUGS (Bayesian inference Using Gibbs Sampling) is a free software

package that uses MCMC estimation to build models. David Spiegelhalter[14], a senior scientist at the Medical Research Council's biostatistics unit in Cambridge, UK, was head of the research team that developed the first BUGS package in the early 1990's. Many of the current WinBUGS programmers (e.g., Andrew Thomas and Dave Lunn) are now based at Imperial College, London, in the department of epidemiology. (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml, cited 14-07-05).

In psychology there are few applications of Bayesian inference but recommendations have been around for decades (Rozeboom, 1960; Edwards, Lindman & Savage, 1963). There are more recent advocates too, but as in medicine, they still constitute a relatively small minority of NHST critics (Gill, 2002; Pruzek, 1997; Rindskopf, 1997; Winkler, 1993).

2.2 Defences of NHST

In medicine and ecology there have been few defences of NHST. Altman (2000a) briefly mentioned some "dissenting voices" in medicine but quickly noted that there have "been very few who have spoken out against the general view that confidence intervals are a much better way to present results than P values." (p. 10). Altman also noted that there is no consensus amongst this small group as to what the failings of CIs are.

In psychology, on the hand, there have been several explicit arguments for the continued use of NHST—some more systematic (Chow, 1996, 1998, 2000) than others (Frick, 1999; Hagen, 1997; Wainer, 1999). In the last ten years, the situation in psychology could accurately be characterised as a 'debate'. However, it is important to recognise that the point of difference is between those who, on the one hand, believe that when interpreted properly and supplement with effect sizes and statistical power, a *p* value provides useful information and NHST reporting should continue with caution (Abelson, 1997; Mulaik, Raju & Harshman, 1997) and, on the other, those who believe the use of NHST has been so damaging as to justify abandoning the procedure altogether (Hunter, 1997; Schmidt, 1996). The serious debate (in as much as it exists) is *not* over whether current practice, with its neglect of statistical power and slavish adherence to the .05 level amongst other flaws, should continue. In other words, there

[14] Spiegelhalter's name may be familiar from discussions of ecology; he has published on Bayesian and information theoretic methods in both disciplines.

is consensus that if used at all, NHST should be supplement with statistical power analyses, statistically non-significant results be interpreted with caution, effect sizes (of some sort) be provided and that statistical significance not be conflated with theoretical or practical importance. Consensus, that is, with the exception of Siu Chow.

2.2.1 Chow's Defence

In 1996 Chow published *Statistical significance: Rationale, validity and utility.* It was not the first time Chow had defended the current practice of NHST (Chow 1988, 1991) but it was perhaps the first time his efforts drew serious attention. In 1998, the journal *Behavioral and Brain Sciences* published a precise (authored by Chow) of his 1996 book, along with an extraordinary number of commentaries (37 in total), all of which were highly critical in one way or another. In fact, Chow's defence even motivated criticisms from other defenders of NHST. Frick (1998) called Chow's defence "too traditional", adding, "one might expect the defenders of NHSTP to ally, but this alliance too would be unnatural" (p.199)

Chow (1996, 1998) offered lengthy discussions on the value of NHST in theory corroboration. An overwhelming proportion of the commentary on Chow's precise objected to his model of theory development and corroboration (Bookstein, 1998; Dar, 1998; Gigerenzer, 1998; Stam & Pasay, 1998; Vicente, 1998). Despite the many philosophical problems with Chow's model of theory development, this is not the particular argument I want to take issue with here. There are two other aspects of Chow's defence of current practice—namely the neglect of statistical power and effect sizes—that are both more tractable and more relevant to criticisms already discussed.

Chow's boldest claim is that statistical power is not relevant to statistical significance, because the two concepts exist at different levels of abstraction. Harris (1998) accurately summarises Chow's position and provides a counter-argument:

> He [Chow] further argues that since power requires a consideration of two distributions while significance testing is based soley on the null distribution, the two cannot be related. This has the same logical status as arguing that factor A has nothing to do with the A x B interaction because a plot of A ignoring B requires a single line, while the A x B interaction requires a different line (set of means for different levels of A) for each level of B (p.203).

Chow claims that statistical power is often mistaken as "the a priori probability of obtaining statistical significance" (p.170). He rightly points out that statistical power is a conditional probability—and his critics quickly point out that most power analysts are aware of this (Lashley, 1998; Mayo 1998). Chow attempts to provide illustrations of null and alternative distributions, but his illustrations are actually of control and experimental distributions. As a result of this error, the arguments that refer to or rely on these illustrations are largely incomprehensible.

Chow's position on effect sizes runs parallel to his position on statistical power. That is, he claims that effect sizes, like power analyses, are not relevant to questions of statistical significance because they belong to a different level of abstraction. Chow admits effect sizes may be useful in a practical context (though makes very serious qualifications about what constitutes such a context) but denies they are of statistical concern. Nester (1998) correctly criticised Chow on this point—arguing that effect sizes are important statistical parameters, and that together with precision information, can provide evidence for or against a hypothesis.

Chow has continued (and been prolific in) his defence of NHST (see Chow 1998, 1999, 2000, 2002). He has also published criticisms of meta-analysis (Chow 1987, 1996). But as I mentioned earlier, the serious debate is not centred on the validity or utility of statistical power or effect sizes, or any other defences of the kind Chow provides. The only substantiative debate is about whether NHST should continue *given that* typical misconceptions are avoided and supplementary information is provided *or* whether its utility is sufficiently limited, and damage cause sufficiently serious, for it to be abandoned altogether. Examples of the position that NHST should be used, given those provisions, are offered in the next section.

2.2.2 Other Defences of NHST

Mulaik, Raju and Harshman (1997) argue that there is a "time and a place for significance testing" (p.65). They claim that reform advocates have confused misinterpretations of statistical significance tests with the tests themselves. They rightly point out these are separate issues. Mulaik et al.'s explicit purpose in their 1997 chapter was to debunk Hunter and Schmidt's arguments (Hunter, 1997; Schmidt, 1996; Schmidt & Hunter, 1997) that NHST should be abandoned altogether. However, they largely failed to provide an argument for why CIs should not replace NHST, which was Hunter

and Schmidt's proposal. They make a number of bold claims in defence of significance tests—for example, that they "are essential to integrating and unifying conceptually the diversity of our observations into concepts of an objective world" (p. 96) and "we cannot get rid of significance tests because they provide us with the criteria by which *provisionally* to distinguish results due to chance variation from results that represent systematic effects in the data available to us" (p. 81)—yet offer little by way of substantial explanation or evidence.

Harris (1997) also offered a case for retaining NHST, albeit in a revised three value logic form (first proposed by Kaiser, 1960). As for why CIs could not fill the vacancy he simply stated: "confidence intervals… can be employed so as to essentially mimic NHST" (p.159). Clearly not alone in this view, similar sentiments were expressed by Abelson (1997): "Under the law of the diffusion of idiocy, every foolish application of significance testing is sooner or later going to be translated into a corresponding foolish practice for confidence limits" (p.130). For others too, defending NHST rapidly becomes less about why *p* values should be retained and more about why CIs are inadequate replacements. These challenges have to be taken seriously, and they raise important empirical questions.

A final reason for retaining NHST was offered by Grayson, Pattison & Robins (1997)—what they refer to as a 'pragmatic' argument. They claim advocates of CIs have often relied too heavily on oversimplified research scenarios to fulfil their rhetorical needs: "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). On this account, they are essentially right. Too often the examples given in which *p* values are easily replaced by effect sizes and CIs are of simple two-group independent *t* tests or similar. In more complicated multivariate designs how to affect this substitution is not always straightforward. Grayson et al. therefore argue for the retention of NHST in these cases, on pragmatic grounds. There is a strong counter-argument, of course, that because both *p* values and CIs rely on the same information that where-ever it is possible to calculate a *p* value, it must also be possible to calculate a CI. The dominance of statistical significance testing over the last half a century has seen methods based on *p* values advance much further than estimation and CI based methods. These are challenges advocates of reform need to face, and, as I mentioned related issues are discussed in

detail in Chapters Nine and Ten. Chow's defences of NHST, on the other hand, appear to offer little of substance to the debate and not discussed again in this thesis.

3
HAS NHST DAMAGED SCIENCE?

Reliance on significance testing retards the growth of cumulative research knowledge. (Schmidt, 1996, p.115).

If NHST is flawed in some or all of the ways outlined in Chapter Two—and it almost certainly is—an obvious question is what impact has it had on the sciences it dominates? In other words, what damage has been done? There are different views on this matter, even amongst strong advocates of reform. Frank Schmidt and Jack Hunter have argued that NHST has caused considerable damage (Hunter, 1997; Hunter & Schmidt, 2004; Schmidt & Hunter, 1997); that it has, as quoted above, "retarded" the progress of science. Paul Meehl was also "afraid it had done quite a lot of damage" (personal communication, August, 2002). Others, however, are less convinced of these claims. Roger Kirk, for example, presents a common scepticism:

> I'm not convinced that there has been major damage as a result of this. I have not seen evidence to that effect. Schmidt's verbal presentation [1996 APA Division 5 presidential address] and written paper [1996, *Psychological Methods*] did not convince me that we have irrevocably damaged by this strategy… One of the things I've discovered is that multiple people discover the same thing. You look at the history of ideas and you see over and over again, ideas are discovered by independent people within the same year or a couple of months. If great ideas do in fact occur to multiple people, at about the same time, and these multiple people do act on those ideas, suppose that one doesn't reach 'significance' and doesn't get published, others will. That is why I'm not convinced that we've been so terribly damaged. If it's a great idea, it will come to lots of people. (personal communication, August, 2001).

Note however the Kirk has set the bar quite high when he says he is not convinced psychology has been "irrevocably damaged".

The degree of damage is difficult to estimate without a systematic historical assessment of scientific research programs since the 1950s. This is certainly beyond the scope of this (or perhaps any) thesis. In this chapter, I instead offer a list of case studies of damage from NHST—collected from psychology, medicine and ecology. Identifying

cases to include is itself challenging, particular in unfamiliar research areas. Most research literatures are inevitably technical and complex, and isolating single factors that may have sent such a program astray is not straightforward. The cases here are consequently those already documented by others. However, they may not have been framed in this way (as damage), and as far as I know, have not been compiled before.

Even though it is not possible to exactly determine the extent of damage, it is possible to predict the ways in which damage would occur. First, we can distinguish between damage done to the progress of a science and damage done to the subjects of study (e.g., people, the environment). Some of the case studies presented here provide evidence of for both kinds of damage (e.g., 'the case of toe-clipping of frogs', 'the case of intravenous streptokinase for acute myocardial infarction'). Others are focused on the first type of damage—damage to the progress of science (e.g., 'the case of theory of situation-specific validity', 'the case of the phenomenon of spontaneous recovery'). Usually cases that lead to damage to subjects themselves entail damage to science.

There are at least four avenues by which NHST could potentially damage the progress of science:

1. Time wasting on searches for illusory moderating variables to explain illusory inconsistencies in the literature (e.g., the case of the theory of situation-specific validity);
2. Time wasting on incorrect, weak or trivial theories—"the tendency of researchers to spend quite a lot of time on feeble theories have negligible verisimilitude" (Paul Meehl, personal communication, August, 2002)—because statistically significant results are interpreted as providing strong evidence in favour of the theory;
3. Giving up on research programs, and potentially true theories or theories with high verisimilitude, because of the inability to produce consistent results (e.g., 'the case of the phenomenon of spontaneous recovery'). This is the fate Meehl (1978) claimed many psychological theories share with old generals;
4. The loss of those research programs that never get started because of their inability to jump the statistical significance hurdle. Cases of this kind are, of course, almost impossible to document, and make a systematic study of damage virtually impossible.

The following case studies provide examples of at least the first three of these avenues, sometimes several at once.

3.1 The Case of the Theory of Situation-Specific Validity

The claim Schmidt made in 1996 was a dramatic one. To requote: "reliance on statistical significance testing…has systematically retarded the growth of cumulative knowledge in psychology" (p.115). Direct evidence for Schmidt's claim came from a series of meta-analyses he, Jack Hunter and others had done throughout the 1970s and early 1980s. In one of their later papers of this series, Schmidt, Hunter and their collaborators explained that they had originally set out to "empirically test one of the orthodox doctrines of personnel psychology: the belief in the situational specificity of employment tests validities." (Pearlman, Schmidt & Hunter, 1980, p. 373). Along the way, they discovered serious methodological problems in the field and invented meta-analysis[15].

The employment tests were professionally developed cognitive ability and aptitude tests designed to predict job performance. The "orthodox doctrine", namely the theory of situational specificity, held that the correlation between test score and job performance did not have general validity: "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 orthodox doctrine also held that this was the case even when jobs appeared to be superficially very similar. For (a fictional) example, a test might be a good predictor of job performance for an information service operator at a telecommunications company in Melbourne, but not for the same company in Sydney. This might seem strange at first, but it is not implausible. It is not difficult to imagine that subtle differences in the clientele, training structure or supervisorial style could plausibly have a profound impact on job performance. A difficult supervisor at one branch might require successful staff at that branch to have better developed conflict resolution skills. A mass of such subtle differences could seriously challenge a predictive test's claim to general validity. Hence, the orthodox doctrine held that the

[15] Schmidt and Hunter developed their own meta-analytic approach to resolve the apparent inconsistencies described here. Their first publication on meta-analysis was in 1977, just a year after Gene Glass' (1976). However, Schmidt and Hunter did not simply apply Glass' method. They had developed their own techniques independently and in parallel with Glass (Hunter & Schmidt, 1990). The development of meta-analysis is discussed further in Chapter Four.

validity of the tests depended on more than just the listed tasks in a given position description—it depended on the cognitive information processing and problem solving demands of the workplace.

How exactly did the theory of situational specificity arise? The belief in situational specificity grew out of the empirical 'fact' that considerable variability was observed from study to study, even when the jobs and/or tests were very similar. The theory was empirically driven; its purpose was to explain the variability, or inconsistency, of empirical results. However, the effect sizes in most studies were not inconsistent. The apparent inconsistency was in the statistical significance of studies. For instance, imagine study 1 found a particular test to be a statistically significant predictor of job performance at location A; in contrast, study 2 found the same test was *not* a statistically significant predictor of job performance at location B. The purpose of the theory of situational specificity was to explain the inconsistency in the statistical significance of empirical results, by generating potential moderating variables. One obvious factor that also explained why one study found a statistically significant result, and another study did not, was the relative statistical power of the studies. But this went unnoticed for several decades.

The theory of situational specificity grew structurally complex, with addition of many potential moderating variables. In fact, the search for such moderating variables became the main business of industrial or organisational psychology for decades despite the fact that the variability that they had been developed to explain was illusory. In their meta-analyses Hunter, Schmidt and their colleagues demonstrated that the difference in allegedly inconsistent results could be exclusively accounted for by the relative statistical power of the studies. The reporting of individual results as 'significant' or 'non-significant' had created the illusion of inconsistency, even though almost all obtained effect sizes were in the same direction.

> …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).

How long did organisational psychology pursue this misdirected theory and its associated research program? In 1981, towards the end of their meta-analysis 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 Theory of Situation-Specific Validity provides us with at least evidence that NHST, as it is typically used with little regard for statistical power and over-reliance on dichotomous decisions, *can* damage the progress of science—that it can lead a research program widely astray. Whether or not the theory itself is actually true is irrelevant to the argument here. The 'damage' is that years of empirical data were seen to support the theory, when in fact they did not. Hunter and Schmidt 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. (1981, p.1128-9).

Schmidt and Hunter (1998) provide a summary of the "practical and theoretical implications of 85 years of research findings" (p. 262).

3.2 The Case of Learned Helplessness and Depression

Martin Seligman was a pioneer in the study of learned helplessness. The phenomenon itself 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 other cages with separate compartments. They were again administered shocks, but could move into the other compartment to escape the shock. 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's conclusion was that they had learned to be helpless. Immediately, Seligman and his colleagues began to wonder what links learned helplessness (or pessimistic explanatory style) might have with depression and illness.

The strong links between explanatory style and depression and illness soon became apparent (Miller & Seligman, 1973; Seligman, 1972). The effects of helplessness on growth of cancerous tumours and death rates were first observed in rats, and later experiments demonstrated the links in human subjects (Hiroto & Seligman, 1975; Miller & Seligman, 1975). Seligman and his colleagues published at least 25 articles on the topic between 1969 and 1977, and 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. The results from Seligman's lab always showed the relationships between explanatory style and depression, but many attempts to replicate results did not—or so it seemed. It took over another decade to sort out the debate in the literature, which was plagued by 'inconsistent' results.

In *Learned Optimism* Seligman tells the story of being contacted by the editor-in-chief of the *Journal of Abnormal Psychology*. The editors of the journal had decided to devote a special issue to the debate over explanatory style and depression. The special issue in 1978 included a critical reformulation of the learned helplessness theory (Abramson, Seligman & Teasdale, 1978) and a several independent criticisms of Seligman's work. As Seligman (1990) explained:

> The special issue of the *Journal of Abnormal Psychology,* 87 (1978) contained… about a dozen other articles, mostly critical of original helplessness theory, and some heated replies and rebuttals. Since that time there have been hundreds of journal articles and scores of doctoral dissertations about explanatory style, learned helplessness, and depression. This massive literature has been controversial, but consensus has emerged that pessimistic explanatory style and depression are robustly related, as the theory predicts (p.294).

The consensus to which Seligman refers at the end of this quotation came from two sources. First, a meta-analysis by Sweeney, Anderson, and Bailey (1986), which 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 individual studies on depression and explanatory style (or 'attributions' as Robins calls them) had an *a priori* power of .8 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).

Again, the only 'inconsistency' in this literature was an artifact of NHST. The misinterpretation of statistically non-significant results produced by underpowered studies caused debate where there should not have been. The damage here is not in the theory being lost for all time, as in the next case I discuss. Obviously, it was not. However, for at least a decade important theoretical developments and clinical interventions based on relationships between learned helplessness, explanatory style, depression and illness were delayed.

3.3 The Case of the Phenomenon of Spontaneous Recovery

In 1997 Rossi wrote: "the history and fate of spontaneous recovery of previously extinguished or unlearned verbal associations provides an enlightening example of real-world artifactual controversy" (p. 178). In short, the story of spontaneous recovery provides yet another example of how the misinterpretation of statistical significance (and non-significance even more so) is responsible for creating illusory inconsistencies in a research literature. In this case, inconsistency and confusion did not lead to the pursuit of false, mediating variables, or to the delay of important clinical applications of a theory, but perhaps even more seriously, the outright dismissal of a real phenomenon.

What is spontaneous recovery? Pavlov's dogs were trained to associate the ringing of the bell with food, and were thus conditioned to salivate at the sound of the bell. Eventually, just the stimulus (the bell) was enough to elicit the response (salivation): The dogs salivated, even in the absence of food, when the bell rang. But if the association was not sufficiently reinforced, it was lost—the bell no longer produced salivation and the behaviour appeared to be unlearned. However, the behaviour could 'spontaneously recover'. A later bell would sometimes elicit the salivatory response. Spontaneous recovery equates to roughly the same phenomenon—only with humans, not dogs; and rather than salivation, it usually related to remembering complex word/number lists.

Of the published studies on spontaneous recovery, less than half found statistically significant results. That is, in less than half of studies was participants' recall of a previously learned world or number list after it had be extinguished by the

learning of future lists statistically significantly better than chance. According to Rossi (1997) "the resulting ambiguity led to the conclusion in most texts and literature reviews that the evidence for spontaneous recovery was not convincing" (p. 179). Rossi reviewed the literature and where possible calculated effect sizes from individual studies. He then calculated an average effect size across the literature (d = 0.39; 95% CI: 0.27 to 0.48). A modest, but typical, effect size for psychological research: certainly not an effect that would usually be given up as uninteresting.

On the basis of that effect size, Rossi calculated the power of each study and averaged it across all studies. For the average effect size, he found average power of .38 (95% CI: 0.33 to 0.43). Strikingly, 43% of studies reported the detection of an effect. *A priori* nobody should have expected more compelling evidence than this for the existence of the phenomenon. And yet it seems they did; this degree of evidence was deemed equivocal.

However, Rossi's interpretation is not without controversy. In a general review of the *What If There Were No Significance Tests?* (in which Rossi's paper was published) Krantz (1999) challenged the claim that poor use of NHST had led to the abandonment of spontaneous recovery. Krantz argued that Rossi's meta-analysis was flawed through its failure to incorporate the theoretical parameters of the spontaneous recovery research programme, in particular, the temporal specifications that were associated with the prediction of the phenomenon. In other words, the meta-analysis pooled studies with different time delays of word recall and Krantz believed this inflated the effect size.

In order to examine the validity of Krantz' claim, it was necessary to investigate the original spontaneous recovery literature. I read this literature hoping to gain insight into the theoretical structure of the field, especially the importance of temporal intervals, so that I could ascertain to what degree a meta-analysis of this research should have taken this into account. What I found was messy—a distinct lack of consensus. Not only was there debate about whether spontaneous recovery existed (as Rossi suggested) and about the temporal course of the phenomenon (as Krantz suggested) but also about what in fact the theory was; whether it was a short-term memory or a long term memory phenomenon; what its causal mechanisms were; what its theoretical basis was and other things besides. These debates hint at the myriad of possible empirical questions being asked. A not very surprising consequence of this was that many experimental designs were used to study the phenomenon. Such heterogeneity of methodology (in this case

different experimental designs for different theoretical purposes) of course provides a serious obstacle to any overarching meta-analysis.

Returning to the specific question of temporal course, Krantz's objection to Rossi's account was: "The prediction was not merely that spontaneous recovery would occur; rather, it included specifics about its detailed temporal course. Attempts to verify this prediction failed. Spontaneous recovery did not occur at the predicted times, but at times much too short to be counted as support for the theory" (1999, p.1380). Rossi, on the other hand, had a different recollection of the theory: "I do not recall the time course being as substantial a concern as Krantz describes." (personal correspondence, April 2000) (It is worth pointing out that Rossi had some particular expertise in this area; the meta-analysis he published in 1997 was a component of his PhD work.)

In an effort to understand these different positions better, I asked Krantz: "Are you saying that if the effect size that Rossi's meta-analysis shows were known at the time, it would have been irrelevant to the debate?" He confirmed:

> That's more or less correct. To show that an effect exists, on the average, over a number of different alternative conditions, is irrelevant, because the goal was to test a theory that predicted the effect, and the effect was not found under the conditions where the theory predicted it (David Krantz, personal correspondence, April 2000).

I could not find a straightforward, generally agreed upon theory of temporal frames within the literature. There were several theories and a host of variations upon the approach to take to them.

There is certainly some evidence that researchers at the time were aware the phenomenon only existed over shorter intervals. For example Postman, Stark and Fraser (1968) acknowledged that: "whatever limited evidence there is for spontaneous recovery in classical conditioning was obtained for the most part over intervals measured in minutes rather than days." (p. 674). In another well known paper on paper on spontaneous recovery from the same year, Keppel concluded by writing:

> A revision of two-factor theory which is more parsimonious with the facts involves a restriction of absolute spontaneous recovery to a relatively short time interval, e.g., a few hours. (1968, pp. 194-195).

Yet I did not find anything which convincingly suggested they were not interested in shorter time periods, or that observing the phenomenon over these intervals did not fit with the theory. Krantz's claim that there was no interest in a general effect may

therefore be too strong. He is perhaps right to criticise Rossi for pooling these results when presumably they could have been sorted by intervals. Indeed, if it could be established that shorter time periods were of interest and supported the theory, and were the more common cases, then pooling across temporal intervals may have served to 'dilute' (rather than inflate) the effect. If at d= 0.39 the effect size is diluted, then this would indeed be something psychologists would care about! This case, with its messiness and controversy, perhaps more than any other discussed here, demonstrates the difficulties involved in attributing 'damage' to a single cause, such as NHST.

3.4 The Case of Intravenous Streptokinase for Acute Myocardial Infarction

Streptokinase is an enzyme that dissolves vascular thrombi, blood clots caused by atherosclerosis. Theory predicted it would benefit acute myocardial patients, since most cardiac arrests are caused by athrosclerosis—a gradual build up of a fat-containing substance in plaques that then rupture forming blood clots on artery walls.

Between 1959 and 1988, 33 randomized clinical trials tested the effectiveness of intravenous streptokinase for treating acute myocardial infarction. In Figure 3.1 below (taken from Hunt, 1997) the left panel shows a conventional meta-analysis, with 95% CIs for each individual study. The effect sizes are odds ratios, so results are statistically non-significant at p<.05 if the 95% CI includes 1. As can be determined from the figure, the vast majority of these trials (26 of 33) showed no statistically significant improvement at p<.05. The remaining trials did show a statistically significant improvement—and often a dramatic one. When described in this way, that is, as merely statistically significant or not, the results appear inconsistent. Yet, from the figure it is immediately obvious that some trials have extremely wide CIs. This indicates the low precision of these studies. Other intervals are very narrow, which should alert us to the fact that the inconsistency in results, in terms of statistical significance, is likely due to varying statistical power.

The right hand panel gives a cumulative meta-analysis, where the results of consecutive trials are combined. The first CI in this panel is based on the combined results of trial one and trial two; the second interval is based on the results of trials one, two and three, and so on. The striking feature of this panel is that the odds ratios and CIs line up on the favourable side of the null line at OR=1. We can see as early as

perhaps the fourth trial (the seventh at the latest) that the cumulative odds ratio is distinctly less than one (about .75), indicating that streptokinase has a positive effect on the treatment of myocardial infarction.

From the point of view of assessing damage, this means that tests of this treatment could have ended in 1973, or even earlier, if results had been properly collated and interpreted. As it was, with dichotomous decision from NHST clouding the issue, testing continued until 1988 and included some 30,039 further participants. Assuming half of those were control group patients, over 15,000 patients were denied treatment that would have been effective for their condition. Of course, this figure does not even begin to account for incalculable number of other patients that would have presented with this condition during the 15 years from 1973 and 1988, when the Food and Drug Administration accepted the effectiveness of the drug.

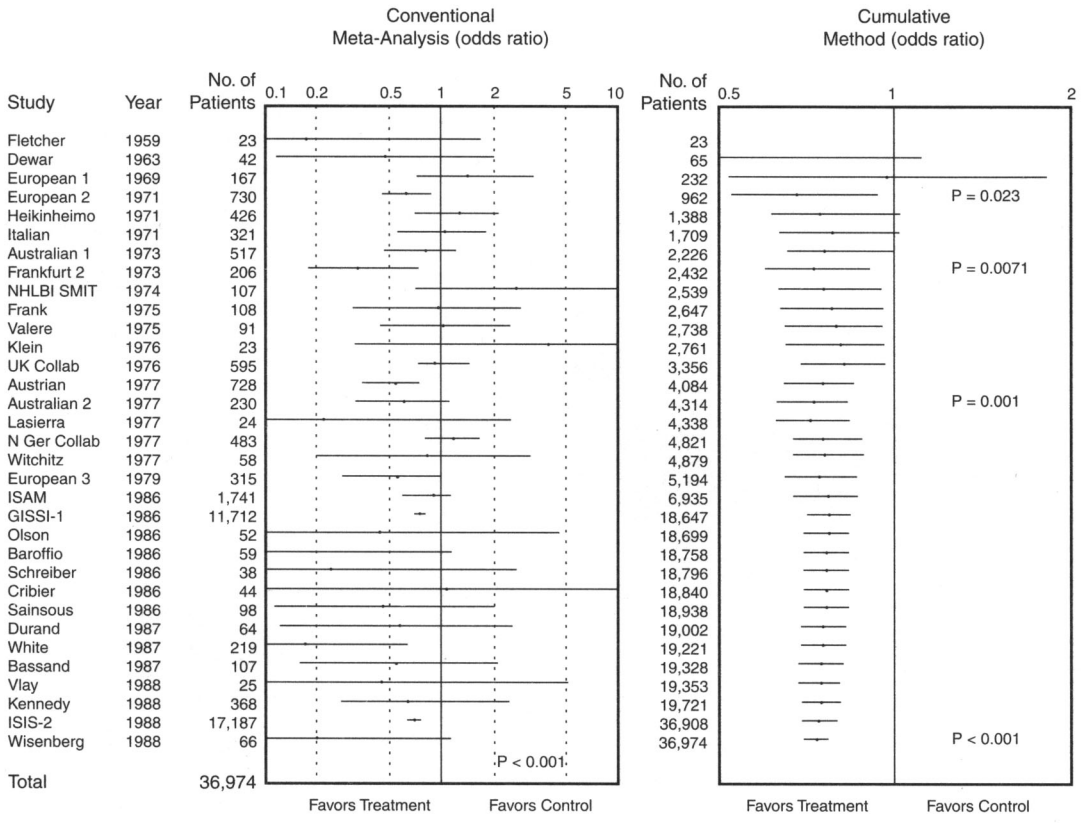


Figure 3.1. Cumulative meta-analysis of trials for Streptokinase as a treatment for Acute Myocardial Infarction. From "The Story of Meta-analysis" by M. Hunt, 1997, citing "Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction" by Lau et al., 1992, *New England Journal of Medicine, 327*, 248-254. Copyright 1997 by Russell Sage Foundation: New York. Copied in reliance of the Australian Copyright Act (1968), Section 40, fair dealing for purpose of research or study.

Clinical Trials in General

It is unlikely streptokinase is an isolated case. Freiman, Chalmers, Smith and Kuebler (1978) demonstrated that therapeutically effective interventions are often missed in clinical trials with too few participants. Almost 95% (67 of 71) of 'negative' (i.e., statistically non-significant) randomised controlled trials in their sample had a statistical power of less than 90% for detecting therapeutic improvements as large as 25%. Even if relaxed to 80%, as is recommended in disciplines like psychology, the proportion of studies in their sample achieving this power would still only be 7%. If the goal was to detect a 50% improvement, still over half the trials had power of less than 90%. Freiman et al. stated: "confidence intervals for the true improvement showed that in 57 of 71 trials a potential 25 per cent improvement was possible and 34 showed a 50 per cent improvement was possible." In other words, true large effects were not ruled out by these studies. Results only failed to achieve statistical significance because of low statistical power. As Freiman et al. explained confusion in the literature was often created because, "in most studies the lack of a difference significant at the 5 per cent level was taken to mean that no clinically meaningful difference existed" (p. 694).

3.5 The Case of Toe-Clipping Frogs

I have already mentioned this case study but it is worth revisiting, especially for the insights it offers into editorial policy. Mark-and-recapture studies are often used in ecology to determine population sizes and survival rates. Such studies require that individual animals are tagged or marked in some way, so that the same individual can be identified upon recapture. With *anurans* (frogs and toads) this can be difficult since their skin is porous and sensitive, and many common marking techniques do not work. The technique used to overcome this is toe-clipping. Certain combinations of toes are removed from individual *anurans* so they can be uniquely identified.

Perhaps not surprisingly, there has been controversy surrounding this method, and several studies have undertaken to asses whether toe-clipping harms *anurans*—whether it increases mortality and consequently decreases return rates. The concern here is two-fold. First, there is concern that a primary assumption of mark-recapture analysis has been violated, that is, that the marking technique itself does not influence the recapture rate. This is concern is foremost about biased population estimates and the progress of science. Second, there is concern that a method employed as part of

conservation studies actually harms that which it seeks to protect. This concern is about killing *anurans*.

Unfortunately, the studies conducted to assess the effect of toe-clipping on *anuran* survival appear to have contradictory results. Parris and McCarthy (2001) reanalysed results from four such apparently contradictory studies (which constituted all the relevant literature). They found that the original four studies had statistical power between 20 and 60% for detecting a decline as large as 40%. (One study in particular only reached 80% power with an expected decline of 75%!) A decline of 40% in this case is an enormous effect size, and could rapidly lead to the loss of an entire population. In McCarthy and Parris' (2004) Bayesian analysis, the population decline in return rate was dependent on the number of toes clipped. Upon removal of a second toe, the return rate decreased by 3.5%; by the eighth toe, the decline was 30%.

This is quite a dramatic case of damage, and one might reasonably expect that those concerned with the conservation of *anurans* would be particularly interested in results that demonstrate their methodology is biasing their population estimates and is harming the subjects of their study. However, Parris and McCarthy's results have not caused the controversy that might be expected, but rather a very different kind of controversy.

They first submitted their manuscript to the *Journal of Herpetology* in late 1998; it was first rejected in February 1999. They subsequently re-wrote the manuscript and submitted it again. It was rejected a second time in September 1999. Some of comments on the manuscript are worth quoting here, as they offer some insight into resistance in recognising and acting on NHST problems. The reviewer (quoted below) makes no mention of the statistical power analysis at all, and has seemingly missed the point that not accounting for the decline in recapture rate caused by the toe-clipping itself is causing bias in population estimates. Rather, they see only an animal liberationist argument, and set themselves—as quantitative scientist—firmly against that position:

> I am not quite certain why but you seem to have an axe to grind with toe-clipping and you have used math to make inferences that could have drastic consequences for the few folks that actually do mark-recapture studies… If this paper is published and if animal-rights activists have this paper to march around prohibiting research, our ability to conduct

> quantitative studies… will be seriously compromised (Reviewer of Parris & McCarthy manuscript, September 8, 1999)

The associate editor agreed with this reviewer, and argued that Parris and McCarthy needed to move from "activism to science" (associate editor of *Journal of Herpetology* in charge of Parris & McCarthy manuscript, September 8, 1999).

Parris and McCarthy then submitted a further revision of their manuscript to a second journal, *Herpetologica*, where the associate editor claimed, somewhat paradoxically, that "it is common knowledge that toe-clipping has adverse effects on survival", adding that their manuscript therefore "provided no real advance" (associated editor of *Herpetologica* in charge of Parris & McCarthy manuscript, May 31, 2000).

Eventually, the somewhat less prestigious journal, *Amphibia-Reptilia* accepted their manuscript in August 2000. In 2004, McCarthy and Parris published a second re-analysis of this data, reporting Bayesian credible intervals for decline in recapture rate after toe-clipping, this time in the renowned *Journal of Applied Ecology*. Following this publication, the issue of toe-clipping become even more controversial!

In September 2004 a letter appeared in *Nature* claiming that there was now convincing evidence that toe-clipping compromised the results of mark-recapture studies. By way of evidence it cited McCarthy and Parris (2004). The letter was written by Lord Robert May (President of the Royal Society 2000-2005, Professor in Zoology, Oxford University, and jointly at Imperial College, London, and until 2000, Chief Scientific Adviser to the UK Government). He directly supported McCarthy and Parris' interpretation, claiming their "statistical study shows convincingly that a technique for marking frogs in ecological field experiments compromises the results. Present practices need a rethink—and not only for practical reasons." (p. 403).

But May's prestige and forthrightness did not intimidate Parris and McCarthy's critics, who, once again missing the point, and replied in their own letter to *Nature*, "several studies have found no negative effects of toe-clipping" (Funk, Donnelly & Lips, 2005, p.193). Funk et al.'s final comment reflected the challenges of activism present in the early manuscript reviews, stating: "we believe it is less ethical to sit back and watch species slip into extinction than it is to use the best available methods to help to conserve them" (p.193). The case of toe-clipping frogs is, I suspect, far from closed!

The potential for damage to the subjects of study is high in applied ecology and conservation biology. When studying small and endangered populations, sample sizes—by definition—will be small and therefore statistical power will be low. The

consequences of misinterpreting non-significant results and declaring 'no impact' is therefore serious: It can result in direct, unanticipated, unacceptable environmental damage. As Taylor and Gerrodette (1993) explained: "The consequences of accepting a false null hypothesis can be acute in conservation biology because endangered populations leave little margin for recovery from incorrect management decisions" (p.489). This type of damage is also evident in the next case.

3.6 The Case of the Northern Spotted Owl

Logging has reduced and fragmented the habitat of the Northern Spotted Owl and population decline is therefore highly plausible. However some surveys, for example Lande (1988), purportedly failed to find a population decline as a result of logging. Taylor and Gerrodette (1993) conducted a power analysis on Lande's results and discovered that the survey had a 64% chance of detecting a decline of 4% per year (considered important in this case), if such a decline existed. Taylor and Gerrodette went on to explain that when environmental variation, not only sampling error, is taken into account the power for detecting a 4% decline was reduced from 64% to just 12%. The first power estimate of 64% was calculated using a single estimate of variance in mortality calculated by pooling data over years. If the variance in mortality over years is taken into account, it becomes even more difficult to distinguish a trend against the background noise and statistical power is further reduced (to 12%). In this case, we can think of 12% and 64% as lower and upper bounds on the true power of detecting a population decline of 4%. Despite this low power range, Lande concluded, on the basis of statistical non-significance, that the population trend observed in the northern spotted owl was not different to that of a stablised population. If 4% per year sounds too small to be of biological importance consider that a decline of this size over 10 years would amount to a loss of one third of the population (Taylor & Gerrodette, 1993).

3.7 Conclusion

As I mentioned in the introduction to Part Two, it is not possible here to demonstrate the extent of damage caused by NHST, or provide evidence that the damage is widespread. Instead, I have outlined pathways by which damage *may* occur and case studies here show that such damage *has* occurred. I would further speculate

that there is nothing strange or unusual about these cases, or the research programs they sit within. If damage can occur in these cases why not in many others?

A question this leads to is whether meta-analysis has the potential to solve such problems. Obviously it was often meta-analysis that exposed these cases in the first place, and in most, finally resolved the issue. Perhaps cases of research programs being led astray for decades, like the case of the theory of situation-specific validity or learned helplessness, are less common in recent years. Although difficult to document, it is quite likely that meta-analysis now 'catches' these cases before so many years are lost. As far as statistical reform efforts go, the relatively widespread uptake of meta-analysis must be counted as the greatest success. Of course, meta-analysis is not retroactive; it can never bring back the years wasted on the cases identified here or others like them. Furthermore, there are some cases to which meta-analysis simply does not apply. For example, it is unlikely the Northern Spotted Owl will survive long enough for a meta-analysis to ever be carried out.

It is curious that despite meta-analysis' relatively widespread endorsement, the very things it presupposes (or at least that would greatly facilitate it)—reporting of effect size and variance information—are still often not reported. Meta-analysis does indeed have the potential to 'clean up' much of our research literature, but this potential will not be fully reached until individual studies routinely report all the information necessary for accurate meta-analyses to be carried out. Furthermore, it can't be good for any science to have the vast majority of its researchers misreading the literature, misunderstanding their own results and mis-designing experiments as a consequence—even if meta-analysis can, in God's own time, clean up the mess.

4
STATISTICAL REFORM IN PSYCHOLOGY

I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories . . . is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology (Meehl, 1978, p. 817).

NHST has been entrenched in psychology since the 1950s. In Chapter One I reviewed some explanations for its popularity amongst psychologists. In Chapter Two, I outlined the many and various criticisms that have been made of NHST. In psychology, such criticisms have been published regularly in well-respected journals for close to half a century. Yet despite cogent arguments against the practice, NHST continues as the dominant methodological procedure.

In one sense it is not surprising that change takes time. This phenomenon is well documented in many histories of science. Thomas Kuhn (1962), for example, provided several examples of slow paradigm shifts: from Ptolemaic to Copernican cosmology; from Newtonian to Einsteinian Physics. Part of the reason such changes are slow, according to Kuhn, is that often an older generation of scientists must retire or die before new theories or methodologies are adopted as mainstream. The new paradigm may attract the new generation of scientists, but for those with well established careers grounded in the old paradigm, the sacrifice change requires may be too great. A lifetime of work will not be abandoned merely because a new theory has come to light.

Bernard Barber described "scientist's resistance to scientific discovery" (1961, p. 596). Barber identified many cases in the history of science where facts contrary to the dominant theory were dismissed. Amongst many others, Barber identifies the following examples of new ideas that were rejected by the scientific community of the time and/or whose uptake was unduly delayed: Ohm and Maxwell (electricity and magnetism); Magendie (chemistry in medicine); Darwin (evolution by natural selection); and Pasteur (fermentation as a biologically process). Barber's conclusion was that resistance to new ideas is inevitable. Perhaps the longevity of NHST, and passive resistance to alternatives, is nothing more than a simple case of this sociology of science phenomenon?

The case of NHST is almost certainly more complicated than the above proposals can account for. First, it is not even clear that what is required is a paradigm

shift—does supplementing a *p* value with a CI constitute a paradigm shift? Second, whilst several extensions and alternatives to NHST have been proposed—increased use of statistical power, employing a three value logic system to replace the dichotomous system of NHST, increased use of graphics, reporting of effect size estimates, reporting of CIs, modelling techniques, meta-analysis, likelihood methods and Bayesian methods—there has never been a convincing consensus as to exactly what paradigm should replace NHST, or indeed over whether NHST should be replaced at all, as opposed to merely supplemented. Kuhn offered convincing evidence that until such consensus is achieved—until scientists are authoritatively directed elsewhere—there will be no revolution. Persistence with reporting NHST, therefore, falls short of being a curiosity, let alone a mystery.

What remains curious about the case of NHST, however, is the overwhelming lack of response to criticism. By 'response' I mean not only change or revolution, but response in the form of convincing defences against the criticisms or active resistance to particular proposals. In Kuhnian terms, there has been no 'crisis', and the existence of a crisis is, of course, essential to precipitating revolution. In Chapter Two I reviewed some 'defences' of NHST, but in fact true defences of current practice (such as Chow's) are rare, possibly limited to Chow himself. There is resistance, obviously, to giving up NHST in psychology—but it is passive. Neil Thomason's experience helps illustrate the point.

A philosopher of science and advocate of statistical reform, Thomason spent just over two years in the mid 1990s touring psychology departments in Australia and the USA. He gave about a dozen talks on the systemic irrationality of continued typical use of NHST. (His talk often had the inflammatory title, 'Experimental Psychology is the most systematically irrational episode in the history of science'.) His talks were never met with defences of current practices, but rather with resentment or apathy. He recalls a range of responses. Some academics told him they were personally affronted—that he had insulted them and their work. Others acknowledged a problem, but resented it being spoken about. For example, an audience member at an MIT talk afterwards said something like: "we don't need someone to tell us stats in psychology are no good, we know that already" (Neil Thomason, personal communication, July 2005). Thomason was only once asked for further information or references; he was never contacted by anyone wishing to follow up on the topic. One person—out of probably a couple of hundred—indicated to him that they would start using power analysis as a consequence

of learning about these issues, and this person (a doctoral student at MIT) had been Thomason's undergraduate student years before.

This is an experience many advocates of reform will identify with. In interviews (described below), several people recalled similar experiences of passive resistance—whether in the form of apathetic responses from audiences at their own talks, or in the dismissal of their questions (about statistical power, precision, effect size etc) at others' talks. Amongst others, Geoff Loftus, Patrick Shrout and Geoff Cumming, noted that their questions and/or advice had largely failed to influence even other academics in their own department. The situations they described were rarely antagonistic or the resistance active; rather they described an atmosphere of straightforward apathy and dismissal of what was perceived by others as 'statistical nit-picking'.

The literature on NHST reflects Thomason's and others experiences. There are literally hundreds—maybe thousands—of articles attacking the typical use of NHST (and/or its more fundamental attributes) in psychology. My own endnote library contains over a 1000 such criticisms and I could easily add a dozen new ones each month. Yet there is only a scattering of defences, and almost none before 1995. Of those, virtually none provide any seriously damaging arguments against the reform case. Yet despite the lack of response or defence, reformers' arguments affect no change. This brand of passive resistance has no place in the models offered by philosophers of science (such as Kuhn or Lakatos). This history is one of a reform movement gaining momentum in the face of extraordinary inertia. The reform movement has grown, criticisms are published increasingly often, institutions such as the APA have intervened and the controversy now has a high profile in psychology and yet there is still virtually no change in the results that are reported in journals or in the textbooks that are used to teach statistics to psychologists.

As we shall see in this chapter, editorial initiatives to reduce emphasis on NHST began in psychology in the 1990s (e.g., Kendall, 1997; Loftus, 1993; Thompson, 1994). Like the individual published critiques before them, they too had little success in reforming practices. Any change achieved by policy has been short lived and largely failed to spread beyond individual journals (c.f. Finch, Cumming, Williams et al., 2004). In 1996 the APA Board of Scientific Affairs appointed a Task Force on Statistical Inference (hereafter TFSI) to investigate the ongoing controversy surrounding NHST and issue guidelines for statistical reporting. To date, these guidelines

(Wilkinson et al., 1999) also appear to have had little impact on statistical reporting in the journals.

Reform events are presented here roughly chronologically, with some exceptions in the 1990s. This chapter ends just prior to the fifth revision of the APA *Publication Manual* in 2001. This particular reform effort warrants a chapter of its own, see Chapter Five.

Interviews and Methods

Predominantly this chapter chronicles the published reform literature from psychology journals. This perhaps sounds similar to Chapter Two, but my intention here is not to rehash specific criticisms of NHST. Here I explore the different themes and emphases that emerged in different decades, and look at interventions and events that are particular to psychology.

This chapter also relies on material from interviews I conducted with various advocates of reform and members of the TFSI. The majority of these interviews were done in the USA, in 2001 and 2002, when I spent a total of about fourth months travelling to visit universities around the county. (A full list of interviewees follows my Introduction.) Of all the people I initially contacted about an interview, only one declined to participate. There were a few others, who agreed, but travel restrictions (particularly after September 2001) prohibited me visiting. I also conducted some interviews with a smaller number of reform advocates in Australia. Occasionally, interviews would require follow up by email correspondence, so some material cited here may be listed as 'correspondence' and dated post main interview.

Most interviews were conducted in a single session, although some were carried out over several days. I asked interviewees various questions, often specific to their own work and role in statistical reform. One question general to all interviews was "what first alerted you to problems with NHST?" I asked this in the hope of uncovering a particular set of articles or a common story. Whilst there was some consensus over the influence of some famous articles, such as Carver (1978), there were few striking commonalities in reformers' initial motivation. In fact, several could scarcely remember how they first heard of the problems. I also asked for reactions to the TFSI guidelines, and to the statistics section of the APA *Publication Manual* (2001). Most had immediate and thorough answers to these questions, and were quite forthcoming.

What such guidelines ought to contain, most interviewees had obviously given serious thought to.

Finally, I asked what explanations they could give for the persistence of NHST, and associated flawed practices, given their own work and other notable reform efforts. This question—which is the central question of this thesis—surprised many of my interviewees. I got few serious answers to this question, with the notable exception of Paul Meehl's response which I discuss later in this thesis (see Chapter Seven).

The other major supplement to published literature in the current chapter was sourced from archival records of the TFSI, housed at APA headquarters in Washington D.C. I spent 2 days at the APA offices reading and copying agendas, minutes and correspondence relating to the TFSI, and meeting with Dr. Sangeeta Panicker, the APA staff liaison to the TFSI.

4.1 From the Beginning: 1950-1970

4.1.1 Technical and Philosophical Criticisms

In 1960 Rozeboom published "The fallacy of Null Hypothesis Testing" in *Psychological Bulletin*. It was a turning point in the criticism of NHST in psychology (Morrison and Henkel, 1970). There had certainly been criticisms of NHST in the psychological literature before (e.g., Chandler, 1957; Lewis & Burke, 1949), but they had been of a more technical, and optimistic, nature. For example, Lewis and Burke (1949) focused on better use and application of chi-square tests; Chandler (1957) on teaching better interpretation of *p* values. Rozeboom, like these earlier critics, was concerned about common misuse and misinterpretations of the tests, but he went further, expressing a broad pessimism about the purpose of NHST and its poor fit with the way science ought to be done.

This style of non-technical, broad criticism appeared slightly earlier in sociology than in psychology. For example, Selvin's (1957) "A critique of tests of significance in survey research" in *American Sociological Review* outlined broad problems in the application of NHST to non-experimental designs, typical of those in sociology (see also Kendall, 1957; Kish, 1959). Despite the parallels in psychological research, and the plausibility of an exchange between such closely related disciplines, it appears unlikely criticisms such as Selvin's were known by early critics in psychology: There

are no citations to Selvin's or others work in their papers (e.g., Bakan, 1966; Lykken, 1968; Meehl, 1967; Nunnally, 1960; Rozeboom, 1960). Morrison and Henkel, who published an anthology of criticisms collected from both disciplines, noted the lack of interdisciplinary exchange in these early articles:

> …like most of the sociologists, the psychologists exhibit little if any awareness of discussions in their sister discipline (1970, p.210).

This trend of little inter-disciplinary exchange has largely continued, with few reform articles citing references from outside their own discipline.[16] Jacob Cohen's (1969) book on statistical power (discussed later) is one of the rare publications to make an impact on several disciplines—it is referenced not only in psychology, but also in (at least) medicine, ecology and economics.

Whilst reformers in psychology may not have systematically looked beyond the literature of their own discipline, they have retained a sense of history about reform within their own field. Even early critics pointed out that their comments were "hardly original" (Bakan, 1966, p. 423). As any reader of reform literature will know, *many* relatively recent critiques begin—somewhat repetitively—by acknowledging the decades of criticisms that preceded their own. For example, Cohen (1994) opens citing Bakan's words: "If it was hardly original in 1966, it can hardly be original now" (p.997).

Alongside the broad philosophical criticisms of 1960's two other traditions of criticisms were developing. One was empirical surveys of the literature, which assessed what statistics were being reported and how frequently various mistakes or misinterpretations were made. The other was direct studies of researchers' misconceptions about NHST, in particular, about the meaning of a *p* value. Cohen's (1962) review of statistical power is an example of the former trend, whereas Tversky and Kaheman's (1971) assessment of the belief in the law of small numbers and Rosenthal and Gatio's (1963) demonstration of the 'cliff effect', illustrate the latter trend.

Cohen's (1962) survey of then current statistical practices provided the first empirical evidence of widespread neglect of statistical power, as mentioned in Chapter Two. For 70 articles published in 1960 issues of the *Journal of Abnormal and Social*

---

[16] Some exceptions exist, such as a recent issue of the *Journal of Socio-economics*, whose editor (Morris Altman) actively sought to provide an interdisciplinary context to the problem by soliciting articles about reform in psychology, sociology, medicine and ecology.

Psychology, the average power of detecting medium effect sizes was .48—barely equivalent to a coin toss' chance of finding real effects. Only for large effects did it approach reasonable standards, at .83. (Cohen's effect sizes are of course arbitrarily defined, and what is small or large in a given discipline may vary. He took a medium effect size of $d=.5$ to be typical of psychological research). Cohen's study was perhaps the earliest attempt, at least in psychology, to take systematic empirical approach to studying NHST use in published literature.

Cohen was also among the first to offer detailed guidance for overcoming shortcomings of NHST practice. In 1969 he published the first edition of the now classic *Statistical Power Analysis for the Behavioral Sciences*, a guide to calculating statistical power for psychological researchers. (The second edition was released in 1988.) He continued to publish on power and effect sizes for almost another three decades: In 1970, a guide to statistical power calculations for one and two sample tests (Cohen, 1970); a few years later a guide to effect sizes for ANOVA (Cohen, 1973), and many other articles, up until his 1994 "The Earth is Round ($p<.05$)".

At the start of the 1970s, Tversky and Kahneman (1971) provided a direct demonstration of researchers' poor understanding of statistical power. Oakes (1986) described the relationship between the Cohen's literature analysis and Tversky and Kahenman's study: "Cohen studied the power of actual research plans; Tversky and Kahneman studied the power of recommended research plans." (p.14).

First, Tversky and Kahneman asked researchers to estimate the number of subjects that should be run in a replication of an original study, where the original study had 40 subjects. The median estimate, from the 77 researchers they surveyed, for the replication was 20 subjects. The follow up question presented researchers with the scenario that the replication failed to find a statistically significant result (the original study reported $p<.05$). Researchers were then asked to advise a recommended course of action. The most common advice, given by 39% (30 of 77) of researchers, was that an explanation should be sought for the difference between the two studies. This advice is indefensible. The fact that the replication study of 20 subjects had a power of only .43, having half the sample size of the original, should have been reason enough to expect inconsistent results (in terms of statistical significance). To make matters worse, Tversky and Kahneman's sample wasn't a random group of psychological researchers. Their participants were all members of the APA Mathematical Psychology Group. If anyone should have had insight into these issues, it was this group!

Rosenthal and Gatio (1963)[17] study, like Tversky and Kahneman's (1971), was in the tradition of directly studying researchers' misconceptions. Rosenthal and Gatio asked 19 academic psychologists to rate their confidence in results based on varying *p* values and found a dramatic drop in researchers' confidence in results as *p* values fell above .05. They identified the 'cliff effect' as evidence of an arbitrary attachment to $p<.05$ and their results demonstrated the stronghold sharp dichotomous decision making had on researchers' interpretation of results. It was this effect that inspired the later and now famous cry from Rosnow and Rosenthal (1989), "surely God loves the .06 nearly as much as the .05" (p. 1277).

4.1.2 Insights from Meta-Analysis

The 1970s was the birth decade of meta-analysis. There are two independent and parallel histories of the development of meta-analysis—one is Gene Glass' work on psychotherapy; the other is situated in the industrial-organisational psychology literature (including the theory of situation-specific validity, discussed in Chapter Three) and was pioneered by Jack Hunter and Frank Schmidt. These are discussed in turn.

Gene Glass was trained in statistics, but had a deep intellectual and personal interest in psychotherapy and was therefore motivated to see the disputes over the effectiveness of psychotherapy resolved (Hunt, 1997). In particular, he was convinced H.J. Eysneck's claims of the ineffectiveness of psychotherapy were wrong. Glass embarked on a systematic review of the psychotherapy literature, and in the process, invented meta-analysis. First, he and Mary Lee Smith separated studies with control and/or comparison groups from those studies that had a treatment group only. The lack of a control group and/or the failure to provide a sufficient report of data (e.g., articles that reported a *p* value only or only the direction of the effect) dropped the sample from potentially thousands of studies of psychotherapy to just 375 (Hunt, 1997). The effect sizes from studies with an adequate experimental design and sufficient data reporting were then standardised and combined. The overall effect size of the combined 375 studies was .68. Glass translated the meaning of this effect in his presidential address to the American Educational Research Association (AERA):

---

[17] Rosenthal and Gatio's (1963) interpretation of some aspects of their results is controversial (see for example Bakan (1967) and Oakes (1986)). However, these criticisms do not bear on the phenomenon of the cliff effect, which is well established (e.g., Rosenthal, 1964).

> [on average] slightly under twenty hours of therapy, by therapists with two-and-a-half years experience…can be expected to move the typical client from the fiftieth to the seventy-fifth percentile of the untreated population. (Glass, 1976, AERA presidential address, cited in Hunt, 1997, p.34).

By combining the results of the available studies, Glass and Smith had demonstrated that psychotherapy worked. They had also demonstrated that although results from individual studies were often not statistically significant, the overall pattern of an important positive effect was undeniable. Statistical significance testing in individual studies, followed by standard 'nose counting' reviews of the literature, were totally inadequate for exposing the pattern because *p* values are confounded by sample size, and low or medium power guarantees that many *p* values are not small. Attention needed to be focussed directly on the magnitude of the effect *across* studies, which is precisely what meta-analysis does. In short, meta-analysis played a crucial role in exposing the weakness of NHST as a technique for accumulating knowledge in science.

Hunter and Schmidt began their investigation into inconsistencies in the personnel selection literature around the same time as Glass's investigation of psychotherapy research. Here they explain their own parallel developments in methodology:

> Unaware of Glass's work, we developed our meta-analysis methods in 1975 and applied them to empirical data sets from personnel selection research. But, instead of submitting our report immediately for publication, we entered it in the James McKeen Cattell Research Design contest sponsored by Division 14 (The Society of Industrial and Organization Psychology) of the American Psychological Association. Although our development and initial applications of meta-analysis won the Cattell award for 1976, a one-year delay in publication resulted (Schmidt & Hunter, 1977). (Hunter & Schmidt, 1990, p.16)

Both Hunter and Schmidt's and Glass' experience with meta-analyses left them highly critical of NHST. Hunter and Schmidt made perhaps the first, and undoubtedly the most direct, argument for incompatibility of NHST and the growth of cumulative knowledge, and have been outspoken on this matter. Glass, on the other hand, has published little by way of direct criticisms of NHST, although he has explained in an interview with Robinson: "I don't want to mislead anyone who might take my silence

as indicative of some displeasure or lack of interest in these issues." (Glass, personal communication with Daniel Robinson, cited in Robinson, 2004, p.26). He quite plainly preferred effect size measures to NHST results[18]: "Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude—not just, does the treatment affect people, but how *much* does it affect them?" (Glass, personal communication with Hunt, cited in Hunt, 1997, p.29-30).

4.1.3 Critical 1970s Publications

Two other major publications mark the 1970s: First, Morrison and Henkel's *The Significance Test Controversy*, to which I have already referred. This collection of published articles from critics in psychology and sociology was the first systematic attempt (and as far as I am aware, the last for the next 30 years) at creating an interdisciplinary discussion of NHST issues. Whilst Morrison and Henkel's collection is often cited in reform literature in psychology, it is far from mainstream reading in psychology and is now long out of print.

Kirk's (1972) *Statistical Issues: A reader for the Behavioral Sciences*, also an edited book of essays, followed shortly after. Many of the essays in Kirk (1972) also covered philosophical aspects of NHST problems. Kirk had hoped the book would be used more widely than it was, and that it would impact on statistics education. Undoubtedly, Morrison and Henkel would have had similar hopes for their own volume. Kirk for one was disappointed in the response to his collection.

> The reviews were quite good. But I had hoped the book would be used more widely than it was…it wasn't adopted very much. People really liked the book, but they didn't assign it to their students. The intention was that students would be reading it. It was designed to be read by

[18] Not surprisingly, Glass is also critical of meta-analytic techniques based on combining *p* values, rather than effect sizes: "It's ridiculous, this business of combining *p* values of studies" (Glass, personal communication with Hunt, cited in Hunt, 1997, p.29). 'This business'—conducting meta-analyses based on *p* values—does occur. However, those who promote such methods often do so only because "it is likely that published… studies will only report p-values" (Guerra, Etzel , Goldstein & Sain, 1999, p.S605). Rosenthal (1984) proposed such a method, known as the Stouffer method, which perhaps lent some credit to the *p* value approach for a while, given Rosenthal's standing in the quantitative psychological community. However, Rosenthal has since promoted an approach based on effect sizes, specifically *r*, and recently proposed a method for extracting an approximate *r* from a *p* value, should *p* be the only statistic reported (Rosenthal & DiMatteo, 2001). Rosenthal and DiMatteo (2001) also explained how the counter null effect size (Rosenthal & Rubin, 1994) is helpful in meta-analysis for overcoming the problem of individual studies which only report *p* values.

> students in their first graduate course, or even in undergraduate. But what I discovered was that the teachers read the book, but they didn't assign it, they didn't require it. So it didn't get the distribution I had hoped. I don't know why they didn't assign it. It fell short of my hope that it would have a large impact. These were important issues to bring to people, and in that sense it failed…which was depressing. (Roger Kirk, personal correspondence, August 2001).

Carver (1978) and Meehl (1978) were amongst the most influential and often cited articles in the reform literature of this decade. In interviews I conducted, several reform advocates cited Carver's article as a turning point in their understanding of NHST problems. For example, it introduced Bruce Thompson to problems with NHST; he remembers the clarity of Carver's writing and that Carver was arguing this point as a well-known and respected academic (Bruce Thompson, personal correspondence, December 2000). For Kirk, despite having already edited *Statistical Issues*, Carver provided a clarification of the problems he was grappling with: "It influenced me so very profoundly…it was a turning point" (Roger Kirk, personal correspondence, August 2001). Yet, beyond motivating a proportion of a small group of reform advocates, Carver seemingly had little impact on researchers' practices.

Similarly, Paul Meehl saw "almost no impact" on researchers' statistical practice of his 1967 paper, nor his 1978 paper—despite receiving 1000 requests for reprints of the latter. "Why I got that many requests, I do not know. It had almost no impact on statistics practice in the journals." (Paul Meehl, personal communication, August 2002). Meehl partially blamed the lack of impact on the journals his articles were published in: "With that journal [*Journal of Consulting and Clinical Psychology, JCCP*], almost the only subscribers would be counselling and clinical people. You wouldn't even get social [psychology] people." (Paul Meehl, personal communication, August 2002). His 1967 article stood even less chance of reaching psychologists. It was published in the journal *Philosophy of Science*. Yet, why the 1978 *JCCP* article failed to impact on at least the counselling and clinical community is difficult to explain. A clinician himself, Meehl was certainly well known and respected in that community.

4.2 Reform after the 1970s

At first it is difficult to understand why major reforms in psychology had not occurred by the end of the 1970s. By this time, there were criticisms of virtually all technical aspects (Chandler, 1957; Cohen, 1969) and philosophical aspects (Carver, 1978; Meehl, 1967, 1978; Rozeboom, 1960) of NHST. Reform literature was amassing (e.g., Binder, 1963; Grant, 1962; Nicewander & Price, 1978; Signorelli, 1974; Wilson, Miller & Lower, 1964). Critics were not merely unknowns publishing in obscure journals. In addition to the well-respected names already mentioned, Tukey (1969) and Cronbach (1975) added their thoughts on NHST to the literature during this time; both articles appeared in *American Psychologist*, the APA's flagship journal.

Further, by the end of the 1970s there had been clear demonstrations of widespread misuse and/or misinterpretation of NHST in the literature (Cohen, 1962; Craig, Eison & Metze, 1976) and direct evidence of researcher's misunderstanding from surveys of psychological researchers (Beauchamp & May, 1964; Rosenthal & Gatio, 1963; Rosenthal, 1964; Tversky & Kahneman, 1971). There was also growing evidence of the practice causing damage (Schmidt, Berner & Hunter, 1973; Schmidt, Hunter & Urry, 1976; also see Chapter Three).

Supplements and alternatives had been recommended and some guidelines for their application published. Cohen (1969, 1970, 1973) had published extensively on statistical power and effect sizes. Others too recommended effect sizes and provided tables for their quick estimation (Friedman, 1968). CIs had been discussed and proposed as an alternative (LaForge, 1967; Rozeboom, 1960, p. 227). Meta-analysis was established (Glass, 1976; Schmidt & Hunter, 1977; Smith & Glass, 1977). There were even advocates of a shift to Bayesian methods (Edwards, Lindman & Savage, 1963; Rozeboom, 1960, p. 227).

Paul Meehl said of the evidence against NHST accumulated by the end of the 1970s, "the matter was settled" (personal correspondence, August 2002). Jacob Cohen, however, offered an explanation for the persistence of NHST. In his later years, seeing NHST still so entrenched in psychology, Cohen became concerned that his own reform efforts were responsible for its lingering nature: "I'm afraid that my long sojourn with power analysis and Neyman-Pearson has served to prolong the sway of NHST and made our reformist educational job more difficult." (Jacob Cohen, letter to Frank Schmidt, 26 September, 1994, sourced from TFSI archives). In an interview with Neil

Thomason, Cohen again expressed regret at his focus on power, suggesting instead that he should have promoted CIs (Jacob Cohen, personal communication with Neil Thomason, 7 November, 1994; cited in Finch, Thomason & Cumming, 2002).

Cohen's reflection on his "long sojourn with power analysis" does perhaps offer some insight into the persistence of NHST in psychology, but it is, of course, only part of the story. There were other forces working against reform during these years. By the 1970s, NHST was firmly entrenched in psychology, used in over 80% of articles (Huberty & Ryan, 2000; Sterling, 1959). In some cases at least, it was mandated in editorial policy (recall Melton's 1962 editorial at the *Journal of Experimental Psychology*).

Textbooks continued to present the anonymous hybrid identified by Gigerenzer (1993), and statistics curricula in psychology were largely untouched by growing criticisms, leaving little hope that the next generation of scientific psychologists would fair any better. Haller and Krauss (2002) recently commented that although problems with NHST are very well known in psychology "there is astoundingly little pedagogical effort to eliminate these misconceptions."(p.3). In 2002, they were referring to recent or current pedagogical efforts—in the 1960s and 1970s, there was even less than 'astoundingly little'.

Institutions, such as the APA, were also unaffected by calls for reform during these decades. There were, from my knowledge of the literature and interviews, no noteworthy symposia, conferences or debates held on the topic during these decades. Criticisms were growing, to be sure, but they were growing by the single contributions of largely isolated individuals. In short, there was little *organised* reform action. For example, as late as 1994, Jacob Cohen had neither met nor corresponded with Robert Rosenthal.[19] (Jacob Cohen, personal communication with Neil Thomason, 7 November 1994). They were, not surprisingly, aware of each other, and admired each other's work, but despite being only a short train ride away (Cohen was at NYC and Rosenthal at Harvard) they had not met, or collaborated, or organised any symposia or forums on the topic.

Given the limited impact of early criticisms, the entrenchment of NHST in editorial policy and textbooks and the lack of an organised reform movement it is perhaps not surprising that similar criticisms of NHST continued through the 1980s

[19] They did meet decades later of course, and became joint co-chairs of the TFSI.

(e.g., Gigerenzer, 1987; Pollard & Richardson, 1987; Rosenthal, 1983; Serlin & Lapsey, 1985; Shaver, 1985; Thompson, 1989a, 1989b). Yet the reform efforts of the 1980s, like those of the decades before them, had little impact on researchers' statistical practice. For example, at the end of the 1980s, Sedlmeier and Gigerenzer (1989) asked "Do studies of statistical power have an effect on the power of studies?" Finding the average statistical power of articles in the *Journal of Abnormal Psychology* virtually unchanged since Cohen's (1962) survey of the same journal, they concluded they did not. A few other notable reform efforts of the 1980s follow.

In 1986, Oakes published *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Oakes' empirical work provided direct evidence of several widespread misconceptions about NHST particularly about the meaning of *p* values (Oakes' survey and results were discussed in Chapter Two). Oakes' book remains extremely well-regarded amongst reformers. For example, after it went out of print, Kenneth Rothman, then at *Epidemiology Resources Inc.*, reprinted it because he considered it such a clear exposition of the important issues (K.J. Rothman, personal correspondence, August 2001).

In 1987, the two volume *Probabilistic Revolution* was published. Volume 2 *Ideas in the Sciences* (Krüger, Gigerenzer & Morgan) included essays about the history of NHST in psychology (Dazinger, 1987; Gigerenzer, 1987a, 1987b). These essays were generally pessimistic about the direction experimental psychology had taken since the dominance of NHST, and what psychology's future would be like if its reign continued. In another collective effort, the journal *Behavioral Assessment* in 1988 ran a special issue on clinical significance, as distinguished from statistical significance. Articles featured in the special issue offered new insights into how to measure the former (Hollon & Flick, 1988; Jacobson & Revenstorf, 1998). Yet still none of these efforts had a detectable impact on reporting in the journals, or what appeared in textbooks or statistics curricula in psychology.

4.3 A Decade of Editorial and Institutional Intervention: The 1990s

It is not clear why researchers continue [with NHST] …The passive acceptance of this state of affairs by editors and reviewers is even more of a mystery. (Cohen, 1992, p.155).

The 1990s saw reform attempts reach a new level. Not only did the number of individual articles criticising NHST and calling for change increase dramatically, but for the first time in psychology, such criticisms became an editorial and institutional concern. Several critics of NHST have proposed that the editors of major journals could be key players in statistical reform. For example Kirk (1996) claimed a change in editorial policy "would cause a chain reaction: statistics teachers would change their courses, textbook authors would revise their statistics books, and journal authors would modify their inference strategies" (p.757). Similarly, Sedlmeier and Gigerenzer (1989) argued: "There is only one force that can effect a change, and that is the same force that helped institutionalise null hypothesis testing as the *sine qua non* for publication, namely, the editors of major journals" (p. 315).

The 1990s saw these theories tested, with several editorial attempts at changing practices in the journals. A number of such attempts are discussed in detail below. The APA's TFSI represented the first official recognition of this problem by a major psychological institution. I start this section with the TFSI, and the events leading to its formation, then follow with editorial interventions in individual journals. Whilst this puts things slightly out of chronological order, it helps highlight the connections between key reform events in this decade.

4.3.1 The APA Task Force on Statistical Inference in Psychology

In 1996 the APA Board of Scientific Affairs (BSA) established a Task Force on Statistical Inference (TFSI). The prestigious group was given a charter to investigate the growing controversy over NHST, in particular whether NHST should be banned from APA journals. Eventually, a sub-group would become responsible for redrafting statistical recommendations in the APA *Publication Manual*. In all, this was a reasonable enough proposal for the APA to approve—but why in 1996? Given the long

history of criticisms, what precipitated the APA's interest at this seemingly arbitrary juncture? To put it simply, they got a letter.

Dr. Albert Bartz, a psychology academic with no previous involvement or particular interest in statistical reform, came across Jacob Cohen's (1994) "The Earth is Round ($p<.05$)" in *American Psychologist*. *American Psychologist* has, of course, a very wide readership. Being the flagship journal of the APA, it is sent automatically to all members. After reading Cohen's article, Bartz contacted the journal to make a proposal. In 1995, after being referred on, he wrote to the BSA. The agenda at the following BSA meeting stated:

> Albert E. Bartz, Ph.D., is requesting that BSA form a task force to look into the problems of null hypothesis significance testing (NHST). (BSA Agenda, November 1995, sourced from TFSI archives).

The story is of course not quite that simple. Bartz's letter did not arrive at BSA unaccompanied. When he first contacted *American Psychologist*, they referred him to Frank Schmidt. Schmidt had just served a term as president of Division 5 (the mathematical and statistical division) of the APA. At the time, he had recently given his presidential address to the division, entitled "Data Analysis Methods and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." (A version of this address was later published in *Psychological Methods* (Schmidt, 1996). In this address, Schmidt argued that reliance on NHST was retarding the growth of cumulative knowledge in psychology. Schmidt sent Bartz a copy of the address, and Bartz included it in his proposal to BSA. Schmidt thereby lent his support, and his name, to the proposal.

In the final package the APA received from Bartz was his own letter to the BSA, as well as: (a) a copy of Cohen's (1994) article, (b) a copy of Schmidt's presidential address; and notably (c) substantial correspondence Schmidt had received as feedback to his address—including letters of support from, among many others, Jacob Cohen, Ronald Carver and William Shadish. The vast majority of these letters were positive and supportive of the proposal that current NHST practices in the journals needed to change. Bartz's ensemble could scarcely have failed to command attention from the BSA, especially given authority of the correspondents Schmidt collected.

BSA rapidly instated a sub-committee to make recommendations on the structure and purpose of the proposed TFSI. The sub-committee immediately suggested Robert Abelson and Robert Rosenthal as co-chairs (memo from BSA

subcommittee to BSA, February 28 1996, TFSI archives). Both agreed, and Jacob Cohen was soon invited as a third co-chair. The three met at Robert Abelson's home in New Haven in May 1996 to devise a list of TFSI committee members. Rosenthal sent the following letter to the BSA describing the charter the three co-chairs had agreed upon:

> We agree that beyond the focused goal of considering the proper (and perhaps less than proper) role of significance testing in the analysis of psychological data, other avenues for the improvement of data analysis need to be kept in mind. However, we add the caveat that changes can only be introduced slowly—and we don't want statistical reform to go the way of the Clinton health plan by proposing too much, too fast." (Rosenthal letter to Alice Eagly, chair of BSA, June 20 1996).

This note emphasised caution and is perhaps the first indication that the TFSI would not be responsible for 'banning' NHST in psychology—a sentiment expressed again, explicitly, in their later publication (Wilkinson et al., 1999). The original list of potential members put to the BSA in November of 1996, which exactly matches the actual list of members during the tenure of the TFSI, was:

> Mark Appelbaum, PhD: editor of new methods journal [*Psychological Methods*]
> Leona Aiken, PhD: experienced methodologist and author
> Gwyneth Boodoo, PhD: multivariate expert
> David Kenny, PhD: experienced data analyst and author
> Helena Kraemer, PhD: biostatistican with broad interests
> Donald Rubin, PhD: statistician
> Bruce Thompson, PhD: editor of several journals
> Howard Wainer, PhD: authority on graphical methods
> Lee Wilkinson, PhD: computer software expert; father of SYSTAT
> In addition, Drs. Lee Cronbach, Paul Meehl, Frederick Mosteller and John Tukey will serve as advisors to the Task Force.
> (Board of Scientific Affairs, November 1-3, 1996, Agenda Item No.4, sourced from TFSI archives, notes in original).

Despite his obvious role in proposing reform and co-developing meta analysis, not to mention in motivating the establishment of the TFSI, Frank Schmidt was not invited to join (nor was his long time co-author, Jack Hunter). Mark Appelbaum, a

TFSI member who would later oversee the redrafting of the statistical guidelines for the *Publication Manual*, explained to me in an interview that "they weren't looking for anyone that radical" (Mark Appelbaum, personal communication, August 2001).

At its first meeting in 1996, the TFSI reiterated Rosenthal's earlier sentiments; they identified their charge as "being much more broadly focused on assessing current practices in the analysis of psychological data, rather than merely evaluating the issue of null hypothesis significance testing and particularly the use of the *p* value." (BSA agenda item No. 13, November 7-9, 1997). The interim TFSI report was published on the APA website (http://www.apa.org/science/tfsi.html, last accessed 19-01-05). At the second meeting of the TFSI, in November 1997, members compiled and analysed the many comments they had received on their interim report. The feedback was extensive. The TFSI archives contain many emails and letters—some were specifically solicited from experts, others were responses to an open invitation for feedback. Most were positive, and offered only small suggestions for improvement of the draft. After the second meeting, the TFSI submitted, by way of Robert Rosenthal, three suggested avenues of disseminating the group's final recommendations:

> (1) relevant portions of the *APA Publication Manual* be revised in accordance with the suggestions of the initial report;
> (2) BSA approve the development of a readings volume which would elaborate and reinforce the recommendations of the Task Force;
> (3) BSA assemble a casebook that would provide examples of exemplary analysis of data as well as examples of less helpful approaches to data analysis. (Draft minutes, BSA meeting 7-9 November, 1997).

BSA rejected the second and third proposals. The second proposal of developing a readings volume was judged not viable: "Members of BSA believe that it would not be particularly effective as a primary means of communication", adding they were "doubtful that a book of classic articles would be widely purchased, because the articles are already available" (Memo to the TFSI from BSA; November 17, 1997). The third proposal of assembling a case book was thought to be too narrow: "If it consisted only of a presentation of cases, it might not be helpful enough to a wide range of psychologists" (Memo to the TFSI from BSA; November 17, 1997).

They did, however, accept the first proposal—to revise the statistics guidelines in the *Publication Manual*. They further recommended that before the *Publication Manual* revision, the TFSI publish an article in *American Psychologist* "as a means for

initiating discussion in the field about changes in current practices of data analysis and reporting." (BSA agenda item No. 10, March 20-22, 1998).

Per the BSA's recommendation, the TFSI drafted an article for *American Psychologist.* This, like their initial 1996 report, was circulated for comment. Most feedback was positive—but not all. The most substantial and critical feedback received by the TFSI was from Paul Meehl. Meehl had been appointed as a senior advisor to the TFSI, and so took particular interest in the drafted article. In the following extended quotation from my interview with him, Meehl explains his correspondence with the TFSI regarding the draft. He described the letters he refers to in the quotation below in detail in this interview. However, I could not locate them in TFSI archive. I found many other letters sent as feedback to the draft, but Meehl's appear not to have survived or, at least, not to have been kept with this other correspondence.

> I'll tell you a story that might interest you. It is a sad commentary on our profession.
>
> The Task Force appointed four outside consultants: Cronbach, Tukey, Mosteller and Meehl. In my letter of acceptance [to be a consultant] I wrote about NHST and the difference between a substantive theory and a statistical hypothesis. I said the 'logical problem of inductive inference is bigger than the mathematical problems being debated, like how you best compute the power for example.'
>
> The first draft [of the TFSI report] had nothing in it of what I had said. So, I wrote another note, and reminded them of my first note. I said 'if what I had to say on this question is all baloney seems to me you might want to tell me what is the matter with it.'
>
> There was no response. The second draft had nothing [related to my comments].
>
> Then, the quasi-final draft arrived, still no reference to anything I had said.

> Finally, in my last letter, I was slightly irritated—I don't have a real fragile ego so I wasn't enraged, but I was hurt—I asked 'I wonder why you appointed expert, outside consultants, if you won't pay any attention to their input.'
>
> Still no response! It is somewhat discourteous: You appoint somebody as an outside advisor and they put in the work. I don't even know whether the chairman of the committee even circulated my stuff.
>
> When I read the final report, most of the things were very obvious and trivial and should have been in there. For example, tell [the reader] whether you've got this population and tell whether people dropped out. Of course, I agree with all that. But on the hardest part of it, the whole problem of inductive inference in this context, what your general view of theory testing is, the philosophical aspects—they were practically missing. You would think that philosophers of science didn't exist! (Paul Meehl, personal correspondence, August, 2002)

I realised after this interview that it was not the first time Meehl had expressed these views of the TFSI. In 1998 he gave a talk upon receipt of the James McKeen Cattell Fellow award at a meeting of the APS in Washington, D.C. In this address, he criticised the APA for being so slow in reacting to criticisms of NHST: "It took 30 years of such unanswered criticism before our sister association, the APA, woke up to the fact that there might be a problem here, and appointed a committee to examine it." (p. 5). Then he directly criticised the efforts of the TFSI itself:

> I don't wish to be invidious, but I am afraid that the APA committee has labored to bring forth a mouse. The report [draft of Wilkinson et al., 1999] reads like a politician's 'blue-ribbon' committee, coming out in favor of motherhood, the flag, and apple pie; and it has no teeth in it. It does not require or forbid anything, including the most irrational current practices. I was one of the four outside experts, named as consultants—the others being Mosteller, Tukey, and Cronbach—and I would be curious to know whether the committee paid as little attention to the other three as they did to me (p. 5).

With Meehl's issues still unacknowledged, the TFSI recommendations were published in August 1999 (Wilkinson et al., 1999). The group's second charge, to revise the

Publication Manual, was underway shortly afterwards. (As I have indicated, Chapter Five examines the *Publication Manual* revisions in detail.)

The TFSI guidelines do not appear to have had great influence on reporting practice in psychology, at least not their specific recommendations to de-emphasis NHST in favour of effect size and interval estimation. Even in 2003 and 2004 issues of leading psychology journals NHST still dominated; CIs were rarely reported (Coulson, Fidler, Cumming, 2005).

4.3.2 The Fourth Edition of APA Publication Manual (1994)

Whilst the TFSI marked the first official institutional recognition of a controversy over NHST, the fourth edition of *Publication Manual*, released in 1994, had attempted to deal with some related issues. The fourth edition was the first to recommend statistical power and the reporting of effect sizes. However, it was not entirely successful in its attempts to clarify issues. For example, the section on statistical power was vague. It recommended power but failed to differentiate *a priori* power from retrospective power and it did not provide examples of reporting practice. This was the recommendation:

> Take seriously the statistical power considerations associated with your tests of hypotheses. Such considerations relate to the likelihood of correctly rejecting the tested hypotheses, given a particular alpha level, effect size, and sample size. *In that regard, you should routinely provide evidence that your study has sufficient power to detect effects of substantive interest*… You should be similarly aware of the role played by sample size in cases in which not rejecting the null hypothesis is desirable. (pp. 16-17, italics added).

The italicised words were almost completely neglected in journal reporting practice over the next decade. In Chapter Five, I will explain that the failure to follow through on recommendations by providing examples of reporting practice and interpretation is still present in the fifth edition. It is perhaps not very surprising then that surveys of journal articles published after the fourth edition found the reporting of statistical power was still extremely low. For example, in 1999—five years after the release of the fourth edition—only 10% of *Journal of Applied Psychology (JAP)* articles reported statistical

power (Finch, Cumming & Thomason, 2001). If the *Manual*'s recommendation had improved this reporting rate at all, it was not to anywhere near the desired level.

It was not only the statistical power recommendation that was ineffectual. The fourth edition's recommendation to report effect sizes also had negligible impact. On the matter of effect sizes, the fourth edition advised:

> Neither of the two types of probability values [predetermined significance level (α) and exact *p* value] reflects the importance (magnitude) of an effect or the strength of a relationship because both probability values depend on sample size. You can estimate the magnitude of the effect… with a number of measures that do not depend on sample size. … You are encouraged to provide effect-size information, although in most cases such measures are readily obtainable whenever the test statistics (e.g., *t* and F) and sample sizes…are reported. (p. 18)

Yet again, no examples were given, and so not surprisingly, the recommendation had little influence on reporting practices in the journals. Kirk (1996) surveyed 1995 volumes of four APA journals[20]. A year after the effect size recommendation of the fourth edition Kirk found 77%, 55%, 12% and 47% of articles (respectively) reported measures of effect magnitude. With the exception of *JAP*[21], effect size reporting was disappointingly uncommon. Vacha-Haase, Nilsson, Reetz, Lance and Thompson (2000) reviewed 10 empirical surveys (including Kirk's) of effect size reporting covering a total of 23 journals. Excluding *JAP*, they concluded "effect sizes have been found to be reported in *between roughly 10 percent ... and 50 percent of articles* … notwithstanding either historical admonitions or the 1994 manual's 'encouragement'" (p. 419, emphasis in original).

Again, if these survey figures entail any improvement in reporting rate as a consequence of the fourth edition, it was far from what might reasonably have been expected from an initiative of this scale. In addition to the primary APA journals, the *Publication Manual* sets the editorial standards for more than 1,000 other journals in psychology, education and related disciplines (APA, 2001). This outcome, then, can

---

[20] *Journal of Applied Psychology*, *Journal of Educational Psychology*, *Journal of Experimental Psychology: Learning and Memory* and *Journal of Personality and Social Psychology*.

[21] See the section on *JAP* in 'Journal Editorial Policies' (later in this chapter) for Kirk's explanation of this difference in reporting rates.

certainly not be attributed to lack of circulation or readership; it must be attributed to the vagaries of the guidelines themselves, and the lack of examples of reporting practice.

4.3.3 APA and APS Symposia on 'Banning NHST'

In 1995 Jack Cohen delivered his "The Earth is round ($p<.05$)" paper to a meeting of the Society of Multi-variate Experimental Psychology (SMEP). Patrick Shrout was at the meeting and recalled:

> He was preaching to the converted. Nobody was going to dispute it with him. The only dispute was over what should be done to change it. During the discussion of that, someone said 'we should ban it', as was done in epidemiology, and we ended up getting these different perspectives that led to the symposia that I organised with Richard Harris for the APA and the APS (personal communication, September, 2001).

As a consequence of Cohen's address to SMEP (and also Schmidt's to APA Division 5 which I mentioned earlier) some SMEP members became involved in an email exchange over NHST problems. The exchange continued for a couple of months (Richard Harris, personal correspondence, November 2001).

The email correspondence eventually led to plans for a public debate. In 1996 both institutions (APA and APS) ran symposia at their annual conventions about whether NHST should be banned from journals. Richard Harris coordinated the APA debate; Patrick Shrout the APS equivalent. Shrout and Harris invited speakers for and against a ban to participate (the same panel of speakers was used at both symposia). Transcripts from the APS debate (the APA debate had included roughly the same talks) were published in *Psychological Science*, with an introduction by Patrick Shrout (Shrout, 1997).

Shrout described the goal of the symposia as having a discussion of NHST issues with a broader audience (personal communication, September 2001). An audience was one thing it had no trouble attracting: "I went an hour early to make sure I could get a seat. In the US, convention centres are huge, the meeting rooms are huge…but there were people standing up, people out in the hall." (Bruce Thompson, personal communication, January 2001).

Shrout had been at the Columbia School of Public Health during the period of Kenneth Rothman's reforms at the *American Journal of Public Health* (see Chapter Six). His

colleagues there started receiving revise-and-submit letters from Rothman in the mid 1980s, requesting removal of statistical significance tests. Shrout recalled: "We were outraged that this happened overnight…These poor epidemiologists who suddenly had the rules changed. Ironically, we were in sympathy with the goal, but we resented the heavy-handedness" (personal correspondence, August 2001). The APA and APS symposia were therefore organised to ensure reform in psychology was instituted differently—and that it included the opportunity for public debate over reform strategies.

The panel was constructed to showcase varying perspectives: Jack Hunter taking the most radical anti-NHST position, and Richard Harris the most committed defence. It was perhaps the first time a defence of NHST received mainstream attention. Siu Chow's work was (and to an extent still is) relatively obscure and virtually no other defences had then been published in high impact psychology journals. There was no obvious candidate to lead the defence team. Harris, who took on the role, had not made any prior contributions to reform literature. It would be too much to suggest the position was invented for the sake of these symposia, but it is not at all clear such defences would have arisen naturally, without being forced by the debate structure of the panel.

Hunter aside, there was little genuine variation in the perspectives offered regardless of the 'side' panel members took. The resulting consensus—that NHST was often misinterpreted, but that it could be a useful tool if properly used—came with apparent ease. This alone is surprising, given the growing controversy in the literature by this time. Hunter was the exception to this consensus, but he had been sidelined in the pre-panel communication. There was considerable concern over how radical his talk would be and more generally what he might 'do' (group email correspondence forwarded by Richard Harris, November 2001). Hunter was known to be a somewhat eccentric character. Perhaps this rendered his position easier for the rest of the panel to dismiss, despite the strength of his arguments?

Hunter's position in the debate itself was also marginalised. How different might the outcome have been with the support of Frank Schmidt? Or Bruce Thompson or Geoff Loftus—who had both by this time taken an editorial stance on the matter? Or, to press the point, with the support of any other respected, serious reform advocate? To be sure, Robert Abelson was a high-profile panel member, but his position on NHST had always been moderate in comparison to Hunter's. Abelson had long argued for

more appropriate use of NHST, but never that its very nature was damaging psychology, or that the technique might be intrinsically flawed. Political speculation aside, the striking absence of other advocates of substantial statistical reform is perhaps sufficient explanation of the apparent emergent consensus.

4.3.4 'What If There Were No Significance Tests?'

SMEP made a decision in 1994 to produce a series of texts, designed to "describe how to apply rigorous multivariate methodology to meaningful research" (Acknowledgements in Harlow, Mulaik & Steiger, 1997). In 1997, *What If There Were No Significance Tests?* became the first title in their Applied Multivariate Applications series. The *What If...?* collection mirrored the APA and APS symposia format, offering for and against arguments about NHST and its alternatives. Also like the APA and APS symposia, the book was in part motivated by Frank Schmidt's 1995 presentation (see acknowledgements in *What If...?*).

4.3.5 Jacob Cohen's 'The Earth is Round ($p<.05$)'

Dr. Bartz was not the only one motivated by Cohen's article. Cohen wrote in a letter to Frank Schmidt that, upon reading 'The Earth is Round ($p<.05$)" another academic, Judy Singer, had also suggested "proposing a statistical advisory committee to the APA Publications Committee to reform editorial practices in APA journals." (J. Cohen, letter to Frank Schmidt, September 26, 1994, sourced from TFSI archives). It is not clear from the archives whether Singer ever sent such a letter to the APA.

Many advocates of reform I interviewed (e.g., Roger Kirk, Bruce Thompson, Paul Meehl, Robert Rosenthal) also attributed the APA's involvement in the controversy (i.e., the TFSI and revision of the *Publication Manual*) to Cohen's (1994) article. To have someone of such high regard, so well known for his work on statistical power, publish a clear and strong argument in a journal of such widespread readership as *American Psychologist*, sent the message that the problems with NHST could no longer be ignored. Patrick Shrout explained:

> He [Cohen] didn't say anything that was terribly new, but he said it in a way that was engaging. He was well known because of his work on regression. He was well known as both a statistically orientated psychologist, but also someone who was mainstream psychology. If you

> look at some of the other papers—for example David Bakan wrote about this in the 1960s—but Bakan was a philosopher of science. When someone like that [Bakan] says something it is not going to be influential. Rozeboom [was another earlier critic] but he strikes people as being kind of a quirky guy. Bakan and Rozeboom made the same arguments. They were coherent, they were logical [but didn't have the impact of Cohen]. So I don't think rational arguments, when you have this kind of social inertia, carry the day. (personal communication, September 10, 2001).

Shrout acknowledged Paul Meehl as someone who should have had more of an impact earlier, "because he's more like Jack [Cohen], more respected in the mainstream." (personal communication, September 10 2001). But as Meehl himself pointed out, his own articles were not published in journals with the readership of *American Psychologist*. (As I mentioned earlier, his 1967 article was published in *Philosophy of Science*; his 1978 article in the prestigious, but still specialised, *Journal of Consulting and Clinical Psychology*.)

There were several other factors in Cohen's favour too. First, his 1994 article came 16 years after Meehl's 1978 article, and criticisms of NHST had amassed in the mean time. Cohen therefore had the weight of all that literature to draw on. Second, and perhaps more importantly, Cohen (1994) was pushing a single point about the misinterpretation of NHST. Meehl (1978), on the other hand, attempted to expose at least 20 problems with experimental psychology—only a couple of which were related specifically to statistical inference (though it seems that the statistical content has attracted the most interest and citations). In other words, Cohen's article was relatively short and focused; Meehl's was long, sometimes difficult, and jumped around between many issues. This alone may explain why Cohen's article sparked institutional intervention and Meehl's did not.

Cohen's 1994 article was, of course, not the first he had written on this issue. In particular, his 1990 "Things I have learned so far" was also published in *American Psychologist* and covered many of the same issues. Yet, it was "The Earth is Round ($p<.05$)" that compelled psychologists to action in a way nothing else published in the previous four decades had. Why? Who can say what is in a name?

4.3.6 Journal Editorial Policies

Geoff Loftus and Memory and Cognition

In 1993, Geoff Loftus, then editor of *Memory and Cognition* requested that authors report results in figures with error bars, rather than tests of statistical significance. His editorial stated:

> 1. By default, data should be conveyed as a figure depicting sample means with associated standard errors and/or, where appropriate, standard deviations.
> 2. More often than not, inspection of such a figure will immediately obviate the necessity of any hypothesis-testing procedures. In such situations, presentation of the usual hypothesis-testing information (F values, p values, etc.) will be discouraged. (Loftus, 1993a, p.3).

Although Loftus said he would "happily consider whatever technique by which an author believes this goal can best be accomplished" (p.3), he emphasised that:

> Over-reliance on the impoverished binary conclusions yielded by the hypothesis-testing procedure has subtly seduced our discipline into insidious conceptual cul-de-sacs that have impeded our vision and stymied our potential… There are often better ways of trying to convey what the data from an experiment are trying to tell us. (p.3).

For Loftus the primary motivation for becoming editor of *Memory and Cognition* was to improve statistical reporting in his field (Geoff Loftus, personal communication, August 2001). He had, since graduate school, been aware of the shortcomings of NHST. His advisor, Richard Atkinson, had discouraged him from using ANOVA, encouraging mathematical modelling instead. During his early career, Loftus battled with journal editors to get his work published without NHST. He often relied on CIs instead. Once, he remembered during my interview, initially having an article accepted without NHST, only to receive requests for the tests after final proofs were approved:

> Everything was done right done to the green form you sign transferring copyright to them. The very last letter [from the editor] accompanying this form said 'I notice you haven't done any hypothesis testing. Please go through and insert the relevant hypothesis tests.' This was from an editor who was well respected, and is still respected a lot. But I was

> infuriated. That was in 1984. (Geoff Loftus, personal communication, August 2001).

In the late 1980s, Loftus was asked to review Gigerenzer and Murray's (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life*. He was pleased to accept the job because "it seemed as if it would provide a public forum for expressing these views" (personal communication, August 2001). His review, "On the tyranny of hypothesis testing", was published in 1991.

Loftus' other articles on this topic (e.g., 1993b, 1996) were generally well received. He remembered: "personal reactions were very positive", but added "the reaction by no means universal" (personal communication, August 2001). The last comment refers most directly to the impact of his 1993 editorial at *Memory and Cognition*.

Finch, Cumming, Williams et al. (2004) surveyed issues of *Memory and Cognition* before, during and after Loftus' editorial and found some increase in articles reporting figures with error bars—from about 7% pre-Loftus to about 41% at the peak of his influence. An increase of 34 percentage points is certainly not nothing: It would be misleading to say Loftus had no effect. As well as an increase in error bars, reliance on NHST used by itself dropped—from 53% to 32%. However, as Finch, Cumming, Williams et al. point out, this meant that at best still less than half of authors were following Loftus' recommendations: "Clearly, Loftus's success was limited. Fully 32% of the articles he accepted were NHST-only and fewer than half (41%) reported any bars." (p. 315). Furthermore, even when error bars were reported, they were rarely used as the basis of interpretation. Statistical reporting may have improved a little, but reasoning about and interpreting results was apparently virtually untouched by the policy change.

Not only was the impact limited, it was also short-lived. After Loftus' editorship term ended in 1997, the proportion of authors following his recommendations dropped off quickly. By 1998-2000, the proportion of NHST-only articles in the journal had risen again to 49% (from 32% under Loftus) and articles including figures with error bars dropped to 27% (41% under Loftus). The two editors immediately after Loftus made no explicit attempt to continue with his policy (Finch, Cumming, Williams et al.).

The change Loftus did achieve failed to transfer to articles that the authors published in other journals. Finch, Cumming, Williams et al. traced the publications of each of the authors in the *Memory and Cognition* survey that followed reformed practices under Loftus and found that in only 22% of cases did use of error bars transfer to the next published article. Clearly Loftus' requests failed to effect most researchers' appreciation of NHST problems. They were merely jumping editorial hurdles and because the hurdles did not exist at other journals, for an overwhelming majority, the practice was soon ignored.

How could it be that an editor so committed to reform achieved only minimal and short-lived impact? It was not due to a lack of support from associate editors. As Finch, Cumming, Williams et al. reported: "Loftus's associate editors were cognizant of his views on data analysis and presentation and generally supported and followed his guidelines; reviewers were aware of Loftus's philosophy and did not deviate substantially from it" (G. R. Loftus, personal communication, June 29, 2000)" (p.317). The resistance Loftus encountered came directly from would-be authors who simply failed to follow his recommendations. He estimates having calculated error bars himself in approximately 100 cases, for authors who failed to provide them (Finch, Cumming, Williams et al.). In my interview with him, Loftus explained:

> A typical example is this. An author fails to reject the null hypothesis of two conditions being different. Suppose, for the sake of argument, they are dealing with 'probability correct' and they have two conditions: one gives a probability correct of 50% and the other gives a probability correct of 65%. They conclude that there is no [statistically] significant difference and then go on to argue as if the two conditions are actually identical. I would calculate the difference plus or minus the standard error and say 'look, the population mean difference may be zero, or 20% in one direction or 40% in the other direction.' I mean its nuts! We shouldn't be publishing this data. [I would tell them] 'Come back when you have something more powerful.' (personal communication, August 2001).

Researchers clearly struggled with the new recommendations. As Loftus explained, "Many people seemed to confuse standard errors with standard deviations… And many people seemed to exhibit deep anxiety at the prospect of abandoning their *p* values." (G. R. Loftus, personal communication with Sue Finch 6 April 2000, cited in

Cumming, Williams et al, p. 317). The resistance from authors somewhat surprised Loftus, although he recognised that NHST was firmly entrenched in psychology and that change would not be easy (Geoff Loftus, personal communication, August 2001). Loftus holds standard statistical software packages partly responsible for maintaining the NHST status quo in so much as the packages "continue to foster the default they are based on" (personal communication, August 2001). As a single editor, attempting to shift practice away from such a firmly entrenched procedure—and still some years before support from institutions such as the APA—it is perhaps remarkable that Loftus' reform had even the minimal impact it did.

Bruce Thompson: Journal of Experimental Education and Educational and Psychological Measurement

Bruce Thompson's interest in problems with NHST began after reading Carver's (1978) article:

> I saw it shortly after it came out and it really struck me because Carver is a very bright and respected scholar and that journal [*Harvard Education Review*] is respected journal. It was an extremely clear and very negative about significance testing" (personal communication, January 2001).

Thompson also listed meta-analysis as a turning point: "Developments like meta-analysis are important, because they change the way people think… it made people aware of effect sizes" (personal communication, January 2001). Thompson has been a strong advocate of effect sizes (e.g., 1996, 1998a, 1999a, 1999b) and more recently of CIs for effect sizes (e.g., Fidler & Thompson, 2001; Thompson, 2002a, 2002b).

From 1993 to 1998 Thompson served on the executive committee of the *Journal of Experimental Education*. To make changes to editorial policy he had to present a strong case to the other editors. "That took time," he remembered, "but ultimately I was able to get them to adopt editorial policies requiring effect sizes." (personal correspondence, November 2004). In 1993 the journal published a special issue on NHST called "The role of statistical significance testing in contemporary analytic practice". The issue included articles by Huberty, 1993; Shaver, 1993; Snyder and Lawson, 1993; and Serlin, 1993. Thompson recalled: "That helped [with getting the policy through]…it made it more reasonable, because we could say 'consistent with the special issue…'" (personal communication, January 2001). What also helped was the 1994 fourth edition of APA *Publication Manual* with its first recommendation for effect

size reporting. Despite the shortcomings of the *Manual's* recommendations, it provided leverage for editors like Thompson to institute more effectual changes (Bruce Thompson, personal communication, January 2001).

In 1997, formal recognition of the journal's editorial policy was made. The journal's guidelines for authors were updated: "Authors are required to report and interpret magnitude-of-effect measures in conjunction with every p value that is reported" (Heldref Foundation, 1997, pp. 95-96). On the contents page of the spring 1997 issue an inserted textbox read: "The Journal now requires authors to report effect sizes with statistical significance tests." (Heldref Foundation, 1997, p.196).

Thompson and Snyder (1997) surveyed 22 articles published in the *Journal of Experimental Education* between 1994-1996, assessing the impact of the journal's 1993 special issue and the APA *Publication Manual's* 1994 encouragement to report effect sizes (the survey predated the formal recognition of policy). They found despite the special section and the *Manual* almost all articles used ambiguous language (e.g., 'highly significant') and only 4 of 22 articles consistently focused on effect sizes, reporting them for each hypothesis test and interpreting them as effect sizes. Another six articles reported some kind of effect size (e.g., correlation coefficient) but did not acknowledge or interpret such statistics as effect sizes; another four reported effect sizes, and acknowledged them as such, for some tests of hypotheses but not others. The results were rather disappointing. Thompson and Snyder (1997) conceded however that their "analysis involved some *JXE* [*Journal of Experimental Education*] studies published only months after the release of the new APA (1994) style manual." (p. 81).

In 1994 Thompson had another chance at instituting reform. He began what we would be nine year term as editor of *Educational and Psychological Measurement (EPM).* His 10 page guidelines for authors warned of many pitfalls associated with interpreting *p* values, explained appropriate interpretation and in particular urged researchers to use the phrase 'statistically significant' rather than the abbreviated ambiguous term 'significant' and to refrain from relying on nil null hypotheses. He required effect size reporting: "Authors reporting statistical significance will be *required* to report and interpret effect sizes." (Thompson, 2004, p. 845, emphasis in original). He also strongly encouraged authors to report results of internal replicability analyses, or better still, external replication studies.

This position was considered extreme by some, because of the mandate on effect sizes. Here Thompson describes his decision:

> As an editor you have to think about what kinds of policies you want…do you want to take a leadership stance or not? The process of editorial writing forces you to think through a position, and whether you are going to advocate something. You have to think through the counterarguments and how people will respond, because you have to anticipate the objections people will have. (Bruce Thompson, personal communication, January 2001).

What impact did Thompson have on statistical reporting in *EPM*? There has been no formal survey (as far as I am aware) of this intervention, but Thompson was confident there had been change:

> I think the biggest impact was on reliability language and reporting. However, (a) more authors did start using effect sizes in validity studies where these were relevant, and (b) used NHST for reliability not at all, or in a more sensible way. (Bruce Thompson, personal correspondence, November 2004).

He believes his policy worked for the following reasons:

> These policies have impacts depending on (a) being clearly articulated and enforced and (b) being in place for a long time. The second condition was met by my editing for 9 years, without a published term of appointment, so people had to assume I wasn't going away. And I rejected outright people who hadn't read and followed the policy. (Bruce Thompson, personal correspondence, November, 2004).

As I mentioned, to my knowledge this journal has not been formally surveyed, but Thompson's reform is perhaps an example of a successful one in psychology. Unfortunately, and not to downplay his considerable efforts, the journal his strictest policy occurred in (*EPM*) is one of relatively low impact in psychology generally (Impact factor=0.756, ISI, 2004) and caters to a very specific audience. It alone was unlikely to have had major broad impact on the discipline, and apparently it didn't.

Kevin Murphy and the Journal of Applied Psychology

In February 1997, Kevin Murphy, the editor of the *Journal of Applied Psychology (JAP)* wrote:

> The *Publication Manual* [4th edition, 1994]…encourages researchers to present effect size estimates…I intend to take this advice to heart. If an

> author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide special justification for why effect sizes are not reported. (p. 4).

JAP had, since the 1980s, reasonably reformist editors. For example Campbell (1982) was critical of over-reliance on *p* values: "Perhaps *p* values are like mosquitoes…no amount of scratching, swatting, or spraying will dislodge them." (p. 693). Guion (1983) explained to would-be authors the importance of reporting effect size measures: "An estimate of effect size not only helps the readers and reviewers evaluate the contribution the article makes, but it may be indispensable for future meta-analyses." (p. 548).

Over the years, journal surveys have documented better than average reporting practice in *JAP*. For example, Chase and Chase (1976) found that the average statistical power in *JAP* was considerably higher than what Cohen (1962) and follow up surveys (Rossi, 1990; Sedlemier & Gigerenzer, 1989) found in the *Journal of Abnormal Psychology*. Kirk (1996) also found higher effect size reporting rates in *JAP* than in the other three APA journals he surveyed. However, Kirk warned:

> Before anyone concludes that authors of articles in the *Journal of Applied Psychology* are more aware of the limitations of null hypothesis significance testing, remember that these authors are more likely to use regression and correlation procedures. Computer packages routinely provide R^2 for these procedures. Authors in the *Journal of Experimental Psychology* [which Kirk also surveyed] are more likely to use analysis of variance of procedures. Computer packages do not routinely provide measures of effect magnitude for these procedures. (p. 754).

The way Kirk discussed standard statistical packages here resembles Loftus' views that standard packages simply promote the default practice. For Loftus that meant fewer figures with error bars; for Kirk, fewer effect sizes.

The claim that *JAP* authors are not necessarily more aware of problems with NHST (despite higher effect size reporting) is supported by Finch, Cumming & Thomason's (2001) survey of NHST reporting practices in the journal. They found few authors reporting exact *p* values—most relied on relative values, e.g., $p<.05$, or 'asterisks' reporting. They also found high levels of ambiguous use of the term 'significant' (that is, authors failing to distinguish whether they meant statistical significance or psychological importance). Even in 1999, only around 40% articles made any mention of the importance or limitations of sample size; just 10% reported

any information related to statistical power. (The reporting of statistical power is still low, despite the average power of studies reported being higher than average). In 150 articles published between 1955 and 1999, they found only four instances of CI reporting.

Philip Kendall and the Journal of Consulting and Clinical Psychology

Within just a month of Murphy's editorial at *JAP*, Philip Kendall, then editor of the *Journal of Consulting and Clinical Psychology (JCCP)*, informed readers:

> Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance (Kendall, 1997, p.3).

For Geoff Loftus at *Memory and Cognition*, the impact of editorial policy was not as dramatic as he would have hoped. However, as I mentioned, Loftus was attempting reform as a single editor in 1993, years before the TFSI had even been conceived, and before even the 1994 *Publication Manual* acknowledged the importance of effect sizes. Unlike Loftus', Kendall's editorial came after: (a) the TFSI was established and their initial report (1996) released; (b) recommendations in the 1994 *Publication Manual* to report effect sizes; and (c) Cohen's 1994 article. Furthermore, it came at the same time as at least one other major journal (*JAP*) was adopting a similar policy. The context might perhaps suggest that Kendall's policy stood a better chance at initiating change. However, like similar policies before, Kendall's had only minimal impact, as we shall see.

In 1999, still under Kendall's editorship, a second reform attempt was made: *JCCP* ran a special section on clinical significance[22]. Articles discussed methods, measures and definitions of clinical significance as well as the associated conceptual difficulties and challenges of assessing clinical significance. Some included 'how to'

---

[22] The *JCCP* special section (1999) is an excellent starting point for advice on clinical significance. Of course, recommendations to differentiate clinical or substantive significance from statistical significance predate Kendall's policy (e.g., Grove & Meehl, 1996; Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991; Kendall & Grove, 1988; Lees & Neufeld, 1994; Meehl, 1954; Rosenthal, 1983). In 1988, a special issue of *Behavioral Assessment* was devoted to defining clinically significant change. There has also been widespread discussion of clinical significance in the medical literature (e.g., Daly, 2000; Lindgren, Wielinski, & Finkelstein, 1994; Luus, Muller, & Meyer, 1989; Manchanda, 1986). More recent articles in psychology include the following: Ogles, Lunnen, & Bonesteel (2001) and Beutler & Moleiro (2001).

guides to calculate various measures (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999; Jacobson, Roberts, Berns, & McGlinchey, 1999; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999) and Kazdin (1999) provided an overview.

Confusion of clinical and statistical significance often manifests itself in ambiguous language. Researchers describe their results as "significant" or "non-significant" without distinguishing whether they are speaking statistically or substantively. Although clinical significance is a matter of judgement, some statistical measures of effect size are especially relevant to clinical research, including the reliable change index (Jacobson et al., 1999; Jacobson & Truax, 1991) and normative comparisons (e.g., Kendall & Grove, 1988; Kendall et al., 1999).

We (Fidler, Cumming, Thomason et al., 2005) surveyed *JCCP* pre and post Kendall's policy and pre and post the special section on clinical significance. In fact, the survey covered four reform events (in bold below):

> Period 1: 59 articles published in 1993, prior to the release of the fourth edition of the APA *Publication Manual* in July 1994;
>
> Period 2: 59 articles from 1996, submitted after **the release of the fourth edition of the APA *Publication Manual*** and published prior to the commencement of Kendall's editorship in 1997;
>
> Period 3: 61 articles from 1998 and 1999, submitted during Kendall's editorship and after **Kendall's 1997 editorial**, and accepted for publication prior to the special section on clinical significance in June 1999; and
>
> Period 4: 60 articles from 2000 and 2001, submitted after publication of **the 1999 special section on clinical significance** and **the TFSI report** (Kendall was still editor during this period).

Our results show some changes that may be responses to calls for statistical reform in *JCCP*. For example, Kendall (1997) asked for "the required effect size" (p. 3), which we took to mean the effect size appropriate in the research situation. For ANOVA main effects this will be the means, or mean differences and/or corresponding units-free measures. There was a notable increase (from 20 to 46%) in the reporting of standardised or units-free effect sizes for ANOVA. Further, in 2000-2001 the percentage of ANOVA articles reporting at least one mean was higher than in earlier periods and the percentage missing at least one mean was lower. Over the first three periods, the percentage of ANOVA articles reporting a mean showed virtually no

change (1993=60%, 1996 and 1998-1999=58%). By 2000-2001, there were about double the number of articles with ANOVA, and the percentage of articles using ANOVA with a mean increased to 82%. The reverse trend was present in reporting rates of ANOVAs *missing* at least one mean. Over the first three periods an average of 63% were missing at least one mean—this was markedly lower in 2000-2001 (23%).

These changes are important and promising indicators of reform. However, for chi-square and *t* tests there was very limited change in effect size reporting. Figure 4.1 shows the percentage of articles reporting standardised effect sizes for ANOVA, chi-square and *t* tests. (We did not code 'raw' effect sizes, such as means and means difference, for chi-square and *t* tests. This is because the coding of such effect sizes was extremely time consuming, difficult and prone to reliability problems, partly due to the way such effect sizes are presented in articles themselves. We therefore limited coding of means and mean differences to ANOVA as it was the most commonly used technique.)

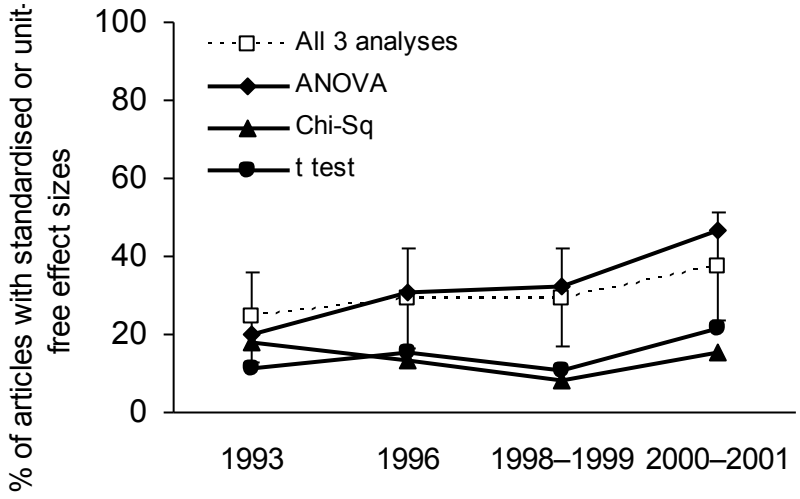


Figure 4.1. Percentage of articles reporting ANOVA, Chi-Square tests and *t* tests which also reported standardised or units-free effect sizes in *Journal of Consulting and Clinical Psychology* between 1993 and 2001. (total n =239; 1993=59, 1996=59, 1998-1999=61, 2000-2001=60). Error bars are 95% CIs.

The frequency of reporting of clinical significance was similar before and after Kendall's policy (36% pre-Kendall and 40% post). Figure 4.2 shows the percentage of clinical significance reporting in 1996 and 2000-2001. However, our coding criteria may have hidden some improvement in the way clinical significance was discussed. Kendall explained that prior to his 1997 editorial and the special section, authors would frequently misuse the term 'clinical' to describe no more than 'statistical' significance

(personal communication, April 9, 2001). Kendall believes this kind of misuse has declined, and that authors are now using more sophisticated measures of clinical significance. Our coding criteria (which consisted primarily of text searches for the terms in Table 4.1) may not always have differentiated between appropriate and inappropriate uses of clinical significance terms. Therefore the percentage of articles published prior to Kendall genuinely considering clinical significance may be less than the 36% we report here. However, there is some evidence to suggest this figure is not a gross overestimate. Dar, Serlin and Omer (1994) reported 30% of *JCCP* articles referred to clinical significance in the 1980s, although they did not make their coding criteria explicit.

Table 4.1.
Criteria for coding presence of clinical significance in *Journal of Consulting and Clinical Psychology* articles. Combinations of words left column and right column; negatives (e.g., non-significant, unimportant) and phrases in the last two rows (and slight variations).

Clinical Significance criteria	
clinically/clinical	relevant/reliance
practically/practical	significance/significant/significant
psychologically/psychological	Meaningful
	important/importance
Reliable Change Index	
Return to normal (functioning)/Indistinguishable from normal	

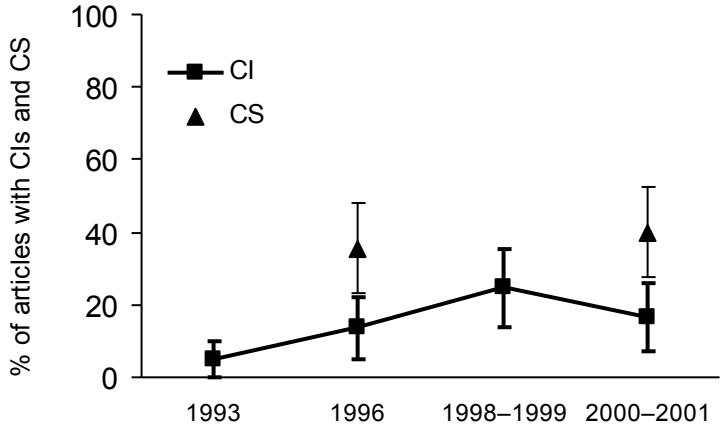


Figure 4.2. Percent of *Journal of Consulting and Clinical Psychology* articles reporting CIs (Period 1-4), and clinical significance (for Periods 2 and 4 only). Bars are 95% CIs.

Even if Kendall is correct about the pre-Kendall figures, our coding of post-Kendall periods shows that, by our criteria, only 40% of articles made any attempt to discuss clinical significance. This is serious. In a major journal dedicated to research on psychotherapy and other interventions, we can assume clinical significance would be relevant to more than 40% of articles.

CIs were infrequently reported (in 17% of 2000-01 articles; see Figure 4.2), even though they were strongly recommended by the TFSI. They were almost never interpreted—only 4 of 239 articles made any reference to the reported CIs—and they were rarely reported in figures, despite this perhaps being an important use (Cumming & Finch, 2005).

In a follow-up to this study, we emailed authors published in the post-Kendall period. Forty-seven of the 214 emails sent were returned undelivered; 62 of the remaining 167 authors replied to the survey, a response rate of 37%. Questions related to awareness of and attitudes towards statistical reform recommendations. A large majority (80%) of authors were aware of at least one of these three reform initiatives: Kendall's editorial, the TFSI report and the *JCCP* special section. Attitudes towards standardised and units-free effect sizes were positive (77% thought they were appropriate to their research), as were attitudes towards CIs (68% thought they were more useful and informative than *p* values). Figure 4.3. shows responses to email survey questions, graphed with 2000-2001 percentage of articles reporting the relevant measure. We acknowledge that respondents to our survey may have been more sympathetic to statistical reform than non-respondents. Further, despite our efforts to choose neutrally-worded questions, respondents may have felt social desirability pressure to give reform-positive responses.

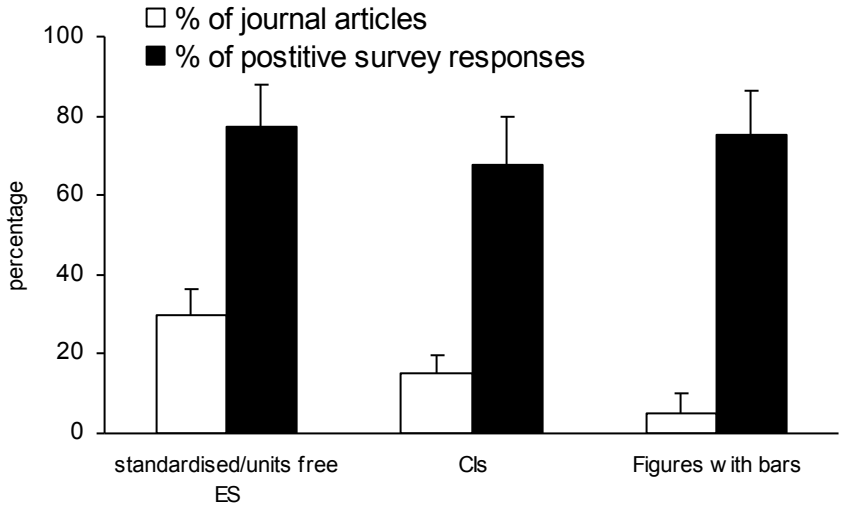


Figure 4.3. Percent of authors' positive responses to survey questions about statistical reform recommendation, and percent of *Journal of Consulting and Clinical Psychology* articles reporting those same measures, in 2000-01. Bars are upper half 95% CIs. The percentage of 2000-2001 ANOVA, chi-square and *t* test articles (n=48) reporting standardised and units-free effect sizes is graphed with responses (n=62) to: "Standardised effect sizes (such as Cohen's *d* for *t* tests, Eta-Squared for ANOVA, Cramer's phi for chi-square) are appropriate to my research." The percentage of all 2000-2001 articles (n=60) reporting CIs appears with responses (n=59) to: "In most cases, it is more useful and informative to report a confidence interval instead of a *p* value." The percentage of 2000-2001 articles with figures that included error bars (n=17) appears with responses (n=60) to: "Graphs that include error bars (i.e., standard error bars or confidence intervals) are preferable to graphs without bars."

From a reform perspective these results offer encouragement, in that positive attitudes may facilitate the changes in practice advocated by reformers. However, there may be hidden difficulties. For example, many respondents considered standardised effect sizes appropriate, and many agreed that CIs are easy to calculate for their data. But CIs for standardised effect sizes require noncentral distributions and iterative procedures, scripts for which have only recently become available for common software and are still relatively obscure (Cumming & Finch, 2001; Fidler & Thompson, 2001; Smithson, 2001, 2002). Similarly, most (80%) respondents disagreed with the claim that figures with error bars are complicated. However, graphing appropriate error bars for mixed or complex designs can, in some cases, result in a complicated figure.

By informing potential authors of desirable statistical practices Kendall took, as editor of *JCCP*, an important and unusual step. Vacha-Haase et al. (2000) reported that only 5 of 50 editorials published between 1990 and 1998 in 28 APA journals addressed statistical reporting practice: Kendall's was one of the most direct. Our results suggest, however, that his policy was, at best, only partly effective in changing the ways that

authors report and interpret their results. Since the rates reported for the 1980s by Dar, Serlin and Omer (1994) there have been some notable improvements: For standardised and units-free effect sizes, from none to 40% (2000-2001 articles in our survey) and for CIs, from none to 17% (2000-2001 in our survey). The reporting of clinical significance in *JCCP* has increased but remains alarmingly low: 30% in the 1980s and 40% in 2000-01. Perhaps some researchers have been slow to pick up on clinical significance because they feel their results are diminished in such a presentation: It is usually much easier to find a statistically significant result than one that is clinically significant.

4.3.7 A New Journal: Psychological Methods

There are of course many individual articles, and probably other collections of articles, from the 1990s that I have not discussed here. The mass of literature on this topic, in psychology alone, grew enormously during this decade. Since many of the articles—though worthy in their own right—repeated the arguments of earlier ones I will not discuss any in particular detail. There is one other event, however, that deserves mention here: a new methodology journal.

In 1992, Roger Kirk became president of Division 5 (Quantitative Psychology) of the APA. During his tenure, he pushed the proposal for a new APA journal for quantitative psychology. He explained the importance of having such a journal: "It gave quantitative psychology more respectability; it gave us a flagship journal" (personal communication, August 2001). The journal was *Psychological Methods*, first published in 1996. By the time the first issue was published, Kirk's term as president was over, and the first editorship of the journal went to the new Division president, Mark Appelbaum.

Since its inception, *Psychological Methods* has published some notable articles on statistical reform issues, including Schmidt (1996) and Nickerson (2000). However, the final product was not exactly what Kirk had envisioned:

> I was a little disappointed, because I wanted the journal to be more tutorial, and speak to the non-specialist as well as the specialist. That is not the direction it took… *American Statistician* does that [i.e., speaks to the non-specialist]. They have a tutorial section which is readable by

> most people that have some training, and it is a chance to upgrade (personal communication, August 2001).

Kirk emphasised that he regards *Psychological Methods*, as it is, "a great journal", and that it did achieve the goal of getting quantitative psychology recognised as a serious discipline. However, he laments the educational opportunity missed:

> We missed that opportunity with *Psychological Methods*. That [education] was my goal and the reason I originally proposed it. It was [supposed to be] a tutorial thing, with a section on critical issues, sections that would inform the reader that was not a specialist. But that is not what happened. I don't know where that idea was lost. I got the letter from the P&C [Publications and Communications] committee saying they approved the journal. And from that point, I was out of the loop; I was no longer president. But I remember my frustration when I saw the first issue—not a single one of my recommendations made it (personal correspondence, August 2001).

There are two things to note about this journal that become important in Chapter Seven, when examining the differences between psychology and medicine's reforms: First that it did not begin publication until the mid 1990s and second, that it was not, in the end, the educational opportunity for spreading reform that it was conceived as.

4.4 The Future of Statistical Reform in Psychology

Decades of cogent criticisms of NHST did little to inspire statistical reform in psychology. Journal editors, as has been suggested (Kirk, 1996; Sedlmeier & Gigerenzer, 1989), may be the key to change, but to date the policies in psychology journals[23] have not been effective in this capacity. Surveys of reporting practice in the journals provide empirical evidence for this claim. These surveys also alert us to the fact that the recommendations in 1994 *Publication Manual* (to report effect sizes and statistical power) have been equally ineffectual. Whilst it is perhaps still too early to predict what change, if any, might follow from the fifth edition of the *Publication*

[23] Currently 24 journals in psychology and related areas have policies on effect size or other reform practices (Hill & Thompson, 2004).

Manual, there is reason to be pessimistic about its influence too, as Chapter Five explains.

By way of an update, Fidler, Cumming, Wilson et al. (2005) surveyed 2003-2004 issues of 10 psychology journals and found overwhelming attachment to NHST and minimal use of suggested alternative methods—for example, less than 10% of articles reported CIs. Couslon, Fidler & Cumming (2005) found similar reporting levels. Nor has psychologists' use or understanding of NHST improved. As discussed in Chapter Two, Haller and Krauss (2002) replicated Oakes' (1986) survey. A remarkable 90% (35 of 39) of academic psychologists agreed with at least one misconception. Even more disturbing is that 80% (24 orf 30) of methodology instructors did too! Haller and Krauss (2002) were astonished by their findings: "Since it can be assumed that the topic of 'significance testing' is addressed frequently during their lectures, this fact is difficult to believe" (p.7).

There is no doubt that, despite the discouraging results from surveys of journal reporting and researchers' misunderstanding, statistical reform in psychology has progressed. The movement has grown and in the last decade or so has drawn the attention of editors and peak institutions. The phenomenon of NHST's persistence in psychology becomes especially curious when we discover that in medicine and epidemiology editorial and institutional reform strategies did, by and large, result in drastic change—at least to reporting practice. Perhaps it is too pessimistic to suggest that reform in psychology will never happen. But it would equally be too optimistic to propose that it is inevitable. As Geoff Cumming put it: "statistical reform in psychology is in the balance" (personal communication, November 2005).

5
THE FIFTH EDITION OF THE APA PUBLICATION MANUAL: WHY ITS STATISTICS RECOMMENDATIONS ARE SO CONTROVERSIAL

They didn't go nearly far enough. I'm very disappointed, very disappointed. (Roger Kirk, personal communication, August 2001).

I don't think anyone was looking for or particularly desired a radical change. I don't think that was the goal. (Mark Appelbaum, personal communication, August 2001).

As mentioned in the previous chapter, the APA *Publication Manual* has been hugely influential in setting the standards of editorial practice in psychology. In addition to the 27 primary APA journals, there are "at least a thousand other journals in psychology, the behavioural sciences, nursing and personnel administration [that] use the *Publication Manual* as their style guide" (APA, 2001, p. xxi). It is "the single text which virtually every psychologist, of whatever sub-speciality, has contact with at some point in their career" (Budge & Katz, 1995, p. 218). Because the *Manual* is so widely known and influential, a revision of its statistics recommendations had been identified as a key step in statistical reform and re-education within psychology. Roger Kirk, for example, argued: "the APA *Publication Manual* and similar manuals are the ultimate change agents" (2001, p. 217). Published in 2001, the fifth edition included more statistical guidelines than any earlier edition.

In the previous chapter, I indicated that the fifth edition seems unlikely to have the expected impact on statistical practice. In this chapter I explore some of the failures of *Manual*. An important question along the way is whether the *Manual's* recommendations reflected the recommendations of the TFSI (Wilkinson et al., 1999). Though the TFSI did not recommend banning NHST, their report was well-received by many reform advocates and has often been cited in articles about improving statistical practice in psychology. The TFSI report unambiguously recommended that the role of NHST in psychological research be de-emphasised:

> It is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual *p* value or, better still, a confidence interval. *Never* use the unfortunate expression "accept the null

> hypothesis." *Always* provide some effect-size estimate when reporting a p value (p. 599, italics added).

The report also recommended practices that reformers had long been advocating, such as presenting effect sizes, CIs, and clear graphics. And it offered a philosophy on analysis:

> Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions... If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories. (p. 598)

Like the TFSI report, the fifth edition includes recommendations to report effect sizes, CIs and graphics. But unlike the TFSI report, it stops short of endorsing statistical reform in any general way. There is minimal acknowledgement of debate and discussion over NHST, but the *Manual* backs away from explaining the importance of the issues or taking any position. Finch, Thomason and Cumming (2002) compared the TFSI report and the recommendations in the *Manual*:

> The TFSI's charter to revise the *Manual*, and the generally detailed and strong TFSI report (1999) both justified expectations that the fifth edition would give an important impetus to reform efforts. The reality, however, is a major disappointment: The fifth edition of the *Manual* (APA, 2001) is very largely, from a reform point of view, a vital opportunity missed (p.839).

Liora Pedhazur Schmelkin was president-elect of Division 5 (Evaluation, Measurement and Statistics) of the APA in 2002 and co-author of *Measurement, Design, and Analysis: An Integrated Approach* (Pedhazur & Schmelkin, 1991). Her opinion of the new *Manual* was similar:

> First of all I was pleased that we had a top-level task force coming out and making some statement... I'm not as pleased with the APA *Manual*, in terms of how they translated some of the recommendations (personal communication, February, 2002).

The *Manual* also failed to follow through on its own recommendations (to include effect sizes, CIs, and statistical power) with examples of how to report these measures. It has, therefore, been interpreted as sending the overall message of "NHST business as usual"

(Finch, Thomason & Cumming, 2002). Some plausibly suggest that specific recommendations are so poorly integrated in the fifth edition that they are unlikely to have any effect on practice. Liora Schmelkin uses the following analogy to describe the lack of follow through:

> I was once told that when you play tennis when you hit the ball you're not supposed to just hit it, you've got to follow through. So even after the ball leaves you, your arm goes all the way up, following it, because that propels it. And I sort of get the feeling with the Task Force report and more so with the *Manual* that they hit the ball, but they didn't follow through. That with a few more sentences, with a little bit more elaboration and some examples you would get a richer sense of what is happening (personal communication, February, 2002).

This failure to follow through on recommendations extends beyond what is contained in the pages of the *Manual* itself. Statistical reform was down played in publicity information targeted at researchers and students. Furthermore, journal editors were not adequately introduced to or encouraged to adopt and promote new recommendations. Indeed, the few sentences on better statistical practices embedded in a large book may even go entirely unnoticed by some editors.

In this chapter I focus primarily on these two types of criticisms: (1) those related to the lack of internal consistency of the *Manual* and its perceived failure to follow through on the TFSI report and to provide systematically integrated examples of newly recommended statistics; and (2) criticisms related to the promotion of the *Manual* and efforts to educate researchers and journal editors about changes relating to statistical reform. This chapter draws heavily on material from the interviews I described in Chapter Four.

5.1 Do the Manual's Examples Correspond to Its Recommendations?

The Manual 'strongly recommends' CIs, yet there are no examples of reporting CIs. They advise taking statistical power 'seriously'. All the examples include p values but never power. (Neil Thomason, personal communication, February, 2002).

5.1.1 Effect Sizes

The recommendation to report effect sizes was stronger in the fifth than previous editions. Additional measures of effect size and strength of association are listed on page 25. The fourth edition "encouraged" (APA, 1994, p. 18) reporting effect sizes; the fifth went further in labelling "failure to report effect sizes" as a "defect in the design and reporting of research" (APA, 2001, p. 5). But the TFSI report included further information that did not appear in the fifth edition. For example, the TFSI report *explained the importance* of effect sizes to future power and meta-analyses:

> We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research (p. 599).

And it drew attention to the importance of interpreting effect size in a practical and theoretical context, and not context-free (p. 599). There are no equivalent statements in the *Manual* even though these issues have long been identified as serious (e.g., difficulties in conducting power and meta-analysis without adequate effect size information were discussed by Rossi (1997); amongst others Cohen (1988) and Rosenthal and Rubin (1982) have discussed the interpretation of effect sizes.)

There is, rightly, concern amongst critics that the recommendations in the *Manual* were undermined by a failure to integrate the new statistics into examples of how to report research results.

> You know how readers are actually going to use it. They are going to try to model examples. They're going to read through, look at some of the major headings and then they are going to go to the examples, thinking "this is a way for me to get my article published. Look at the way they've

> done it here." And they are not going to have what they need to follow in the *Manual* (Liora Schmelkin, personal communication, February, 2002).

Chapter One of the *Manual* covers the "Content and Organisation of a Manuscript" (APA, 2001, p. 3). Within this chapter, section "1.10 Results" (pp. 20-26) deals exclusively with reporting data and analysis. A strong recommendation to report effect sizes is made in this section:

> For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section (p. 25).

Yet this recommendation is not accompanied by any examples of how to report the effect size measures listed. In contrast, recommendations on using NHST made in the same section are accompanied by examples of how to report *p* values. In interviews I conducted, this was repeatedly identified as a failure to systematically follow through on the *Manual's* own recommendations.

In Chapter Three of the *Manual*, a sample ANOVA table has been updated to include a column for effect size, in this case η^2 (p. 162). However, the instructions for constructing an ANOVA table (p. 160) did not mention the effect size column, why it has been included, or how to interpret it. Whilst the sample table has been updated, the accompanying text is the same as that in the fourth edition. The failure to update the text associated with this table provides critics with further evidence that the *Manual* has not integrated reform-related changes.

Towards the end of the *Manual* an example of a complete manuscript is presented (pp. 306-320). Effect sizes are not reported in the results section of this sample manuscript, but *p* values are. Schmelkin again identifies the failure to include an example of the recommendation:

> Now that [the example manuscript starting p. 306 of the *Manual*] is primarily for type setting purposes. But there again, nowhere do you see the words ‘effect size’ (Liora Schmelkin, personal communication, February, 2002).

5.1.2 Confidence Intervals

The recommendation to report CIs is made in section "1.10. Results." The introductory remarks of the TFSI are reflected in *Manual's* statement that CIs "are, in

general the best reporting strategy. The use of confidence intervals is therefore strongly recommended" (p. 22). However, the TFSI report's comments on interval estimates continued with material not included in the *Manual*. For example, the TFSI stressed the importance on comparing CIs across experiments and not against the criterion of whether the intervals subsumed zero. In other words, the TFSI offered advice on using CIs to think meta-analytically: "Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies" (p. 599).

The TFSI report also warned against the "common mistake of assuming a parameter is contained in a confidence interval" (p. 599). The *Manual*, on the other hand, does not alert researchers to this common error. As in the case of effect sizes, there are no examples of how to report CIs. Furthermore, there is no heading for the section on CIs, to draw attention to the new addition.

Michael Smithson (reform advocate and author of *Statistics with Confidence*, 2000) noted the absence of CIs in examples and a lack advice on how to calculate CIs for different parameters. Smithson (2001) explained that in some cases CIs for standardised effect sizes and variance accounted for measures require non-central intervals. He noted in our correspondence that the *Manual* does not recognise this:

> There is no advice or examples [in the *Manual*] concerning how to report CIs or power. For instance, not all CIs are symmetrical so the "estimate plus-or-minus half-width" notation is not always appropriate, but quite a few researchers still are not aware of this (personal correspondence, February, 2002).

5.1.3 Null Hypothesis Significance Testing

The TFSI report made a strong statement about the limited usefulness of NHST, and in particular, of "dichotomous accept–reject" decisions (p. 599). The *Manual*, on the other hand, took no official position: "It is not the role of the *Publication Manual* to resolve these issues" over NHST (pp. 21-22). Yet a strong emphasis on *p* values (and the virtual absence of effect sizes and CIs) in examples of reporting practice betrays the *Manual's* claim of neutrality. For example, in section "1.10 Results" there are three examples of reporting practice. All three contain *p* values; none include either effect

sizes or CIs. In other chapters of the *Manual*, *p* values and statistical significance also overwhelmingly dominate.

Some reformers have further criticised the *Manual's* recommendations for reporting statistical significance, calling them "confusing", "inconsistent" and even "sloppy". "Table Example 7. Sample ANOVA Table" is commonly offered in support of these claims—and is indeed compelling evidence. The table has been updated from the fourth edition so that it now includes a column of exact *p* values. Yet, the instructions for constructing an ANOVA table advise that researchers "avoid columns of probability values" (p. 160)! Furthermore, the table also has a column of *F* values with accompanying asterisks with probability footnotes explaining the asterisks: "*$p < .05$, **$p < .01$" (p. 162). These instructions have not been updated from the fourth edition, when the table did not contain a column of exact probability values!

Asterisks and probability footnotes are entirely redundant in light of exact *p* values. The *Manual* offers no explanation on why both are necessary or how they are to be interpreted. For Robert Rosenthal, the inclusion of asterisks and probability footnotes was one of the most disappointing features of the fifth edition (personal communication, August 2002). As co-chair of the TFSI, he recommended that these items be removed when draft recommendations were sent to him for review. His advice was clearly not followed. In my interview with him he added that in the end he "had little to do with the *Manual*—that committee was separate from the rest of the Task Force" (personal communication, August 2002).

5.1.4 Statistical Power

Section "1.10 Results" also includes a recommendation, first made in the fourth edition, to take statistical power "seriously" (2001, p. 24). The recommendation in the fifth edition has not been updated to include any of the TFSI statements. Absent from the *Manual*, for example, is this explanation from the TFSI report:

> Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size (p. 596).

The *Manual* does not distinguish between *a priori* and *a posteriori* power calculations. Consequently, critics rightly point out: "it isn't clear which the recommendation on page 24 refers to" (Michael Smithson, personal communication, February, 2002). Furthermore:

> ...no recommendations refer to the possibility of including power (and CI width) considerations in describing the study design (e.g., sample size determination) in the Method or Background sections (Michael Smithson, personal communication, February, 2002).

The TFSI report encouraged calculating a range of power analyses, to see how power estimates change for different effect sizes and alpha levels (p. 596). This option is not mentioned in the *Manual,* much less advised. Finally, and perhaps most tellingly, despite an abundance of NHST examples, the *Manual* does not include any examples of reporting statistical power. Nor do its examples follow through on the TFSI recommendation to let "confidence intervals replace calculated power in describing results" (p. 596).

5.1.5 Graphical Representation of Data

There are several inconsistencies between the TFSI report and the *Manual* concerning the use of figures. For example, the TFSI report encouraged the use of well-drawn, simple figures that "attract the reader's eye and help convey global results" (p. 601). The TFSI report also explained: "well-drawn figures need not sacrifice precision" (p. 601). The *Manual*, on the other hand, claims that figures "are not intended to be as precise as tables" (p. 21) and later, that tables "are often preferred for the presentation of quantitative data in archival journals because they provide exact information" (p. 176).

Some of the *Manual's* statements about figures are more positive: "Well-designed figures can covey a memorable image of the overall patterns of results. They also can be the best way to reveal what the reader is not expecting" (p. 177). But, like the fourth edition (1994, p. 15), the fifth edition also discourages their use on the grounds of expense: "figures are more expensive than tables to reproduce" (2001, p. 21). This statement has not been removed, even though with technology advances it is not clear that reproducing figures still incurs substantial extra expense. The *Manual* recommends against "repeating the same data in several places" (p. 21) whereas the

TFSI report explained: "Because individuals have different preferences for processing complex information, it often helps to provide both tables and figures" (p. 601).

The TFSI report also recommended "graphical representations of interval estimates whenever possible" (p. 601). There is no equivalent statement on interval estimates (or error bars) in the *Manual*. However, there are example figures that demonstrate the use of standard error bars. Unfortunately, these have warranted severe criticism. There are three figures with error bars in section "3.77 Types of Figures, Chapter Three." Two show mixed designs with one across-subjects independent variable and one repeated-measures independent variable (pp. 180, 182). Cumming and Finch (2005) explained that the error bars shown would be relevant to some comparisons of means shown in the figures, but virtually irrelevant to any within-subjects comparisons. Geoff Cumming is particularly concerned that no mention of this issue was made in the *Manual*, and the captions of the figures did not make clear that each figure showed a mixed design: “Showing a single interval with such a general label betrays deep misunderstanding” (personal communication, February, 2002). Cumming and Finch (2005) made the following comments on a figure on page 181 of the *Manual*:

> The problem is illustrated by a figure in the Publication Manual (APA, 2001, p. 181), which shows the means for a two-way design with one within-subjects IV. A line segment is shown, with the notation “If a difference is this big, it is significant at the .05 level.” The problem is that different differences need to be specified for each of the two main effects, for simple main effects on either IV, and for any other contrast or interaction of interest. Showing a single interval with such a general notation cannot be correct (p. 179).

5.2 Mandates, Recommendations and Philosophies

The *Publication Manual's* philosophy is restated in the preface of the fifth edition. This paragraph is described as "aptly" characterising the new Manual: "The *Publication Manual* presents explicit style requirements but acknowledges that alternatives are sometimes necessary; authors should balance the rules of the *Publication Manual* with good judgement" (2001, p. xx). The opening remarks in section "1.10 Results" highlight the decision to not offer explicit requirements about statistics:

> The inclusion of a particular approach should not be interpreted as an endorsement of that approach or as a lack of endorsement of some alternative approach (p. 22).

Thompson (1999) had criticised the fourth edition of the *Manual* for being stringent only in regards to the trivial:

> To present an "encouragement" [to report effect sizes] in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't" (p. 162).

Being a TFSI member, he drafted a detailed, three page proposal for the effect size section in the fifth edition, which remains posted on his webpage (below). His proposal made direct reference to the TFSI report and made effect sizes a *requirement*: "Consequently, this edition of the *Publication Manual* incorporates as a requirement, 'Always provide some effect-size estimate when reporting a *p* value' (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599)" (Thompson, (http://www.coe.tamu.edu/ ~bthompson/apaeffec.htm, last cited 10-10-05).

Thompson's proposal to require effect sizes was not adopted. Mark Appelbaum explained that it was agreed early in the drafting process that no "official position" on NHST or statistical reform would be taken in the *Publication Manual* and no particular practices would be banned or mandated (Mark Appelbaum, personal communication, August, 2001). This is, of course, made explicit in "1.10. Results":

> The field of psychology is not of a single mind on a number of issues surrounding the conduct and reporting of what is commonly known as *null hypothesis significance testing*... It is not the role of the *Publication Manual* to resolve these issues (pp. 21-22).

On inconsistencies between the TFSI report and the *Manual* Appelbaum explained the TFSI report was a "statement of principle" which involved very little debate or disagreement among members over substantial issues:

> Leland Wilkinson wrote the article... and we all had a chance to participate in that... It was an onerous task, no one else agreed to do it... I'm sure if I had done it the emphasis might have been a little different, if Leona [Aiken] had done it, if Bob [Rosenthal] had done it. There wasn't a lot of [disagreement]... I think I saw all the feedback, it was more matters

> of taste and style (Mark Appelbaum, personal communication, August, 2001).

The *Publication Manual*, on the other hand, was a different sort of challenge:

> The harder issues came when we were doing the *Publication Manual*, because that got down to some nitty-gritty detail and started becoming a little more prescriptive. Any time you think of one thing to do, someone would come up with an exception. And they'd say "well yeah, it is a good point" and so that was a little more difficult (Mark Appelbaum, personal communication, August, 2001).

Leona Aiken, member of the TFSI subcommittee that drafted the statistics recommendations for the Manual, explained it would have been "irresponsible" to explicitly require practices such as effect size reporting, because of situations where it might not be possible to calculate such a measure (Leona Aiken, personal communication, August, 2001). Aiken's point would not, of course, have precluded a statement of requiring effect sizes 'where-ever possible or practical'!

In the preface to the fifth edition, the debates and compromises involved in producing the recommendations for the *Manual* are noted:

> Mark Appelbaum and his colleagues on the Statistics Task Force (Leona S. Aiken, Joel R. Levin, Robert S. Rosenthal, and Howard Wainer) had a particularly difficult assignment. Although not always in agreement on the specifics, the task force did agree on the need to provide some additional assistance to authors in dealing with statistical representations in manuscripts (p. xx).

Several critics of the *Manual* have been sympathetic towards the decision to be non-prescriptive, but see no reason why this should have excluded following through with examples and explanations. For example, Liora Schmelkin commented:

> I understand the controversy about being prescriptive and saying this is exactly how you should do it. I'm not suggesting that necessarily that should or shouldn't have been the case. But this was a good opportunity to give examples that would be modelling the behaviour we want to encourage. And I really don't find that here (personal communication, February, 2002).

Roger Kirk also agreed with the decision not to mandate but he had hoped the *Manual* would include more explanation of why the recommended statistics matter, and would

warn researchers about common mistakes. In short, he had hoped that the latest Manual would be somewhat more educational:

> It is not forceful enough and it doesn't do enough education for one thing. It doesn't tell them what to do better. It should be more educational, it should be "why they need to do this"... and it offers only a small amount. I expected to see pages and pages on this. So it fell far short of what I hoped would be in there (Roger Kirk, personal communication, August, 2001).

A further educational concern was raised by Sue Finch:

> The examples in the *Manual* are used explicitly in undergraduate courses to teach the new generation of psychologists how to meet APA requirements for reporting statistics. The *Manual* may be intended to be non-prescriptive but it is the defacto authority. Undergraduate students are rewarded for reproducing the model examples in the APA *Manual* in their practice reports. The examples may matter more than the words" (Sue Finch, personal communication, September 2005).

In many other instances the TFSI report also offers considerably more in the way of explanation and education than the *Manual* itself. For example, the TFSI report discussed the importance of comparing and combining results across studies. The report explained that this could be done more easily when effect sizes and CIs are reported in each individual study, so that all researchers had to do was pull these results out of prior reports, rather than face the arduous and sometimes impossible task of themselves computing effect sizes or intervals from previous published research. As I have previously mentioned, the TFSI report discussed the role of effect sizes and CIs in meta-analyses or power analyses; the *Manual* does not. Nor does it otherwise discuss what statistics need to be included in a paper to ensure there is adequate information for future meta-analyses. Also absent from the Manual is any discussion of *why* and *how* "meta-analytic thinking" (Cumming & Finch, 2001) is so important to the growth of cumulative knowledge in the discipline.

Roger Kirk noted the lack of detail given in the fourth edition's statistical guidelines, which he claimed did not match the high level of detail in other areas. Just prior to the publication of the fifth edition Kirk (2001) noted:

> If the 1994 edition of the APA manual can tell authors what to capitalize, how to reduce bias in language, when to use a semicolon, how to

> abbreviate states and territories, and principles for arranging entries in a reference list, surely the next edition can provide detailed guidance about good statistical practices (2001, p. 217).

His criticism applies equally to fifth edition (Roger Kirk, personal communication, August, 2001).

Geoff Cumming similarly identified a difference in the type of guidelines the *Manual* offers on writing and communication and the type of guidelines it offers on statistics. His comment has a slightly different slant to Kirk's: Cumming's concern here is over broad, general advice on what good practice is.

> There's a point of principle here: Should the Manual give such broad advice about data analysis, or simply lay down how particular types of statistics should be reported? There's probably even a precedent. Doesn't the manual near the start give broad, general advice about research communication and the writing of research reports? So why not also for statistics? (personal communication, February, 2002).

Further comments from Cumming highlight still other TFSI recommendations that the *Manual* did not follow through on:

> The Task Force report is impressively broad. It talks about a philosophy of data analysis, even of doing research. It advocates a sophisticated, quantitative approach to theory building, plus use of data for theory testing. Also exploratory data analysis, and the use of minimally sufficient data analysis: Never use a complex statistical technique merely for its own sake! These sort of broad recommendations go way beyond the traditional use of NHST for dichotomous decision making and, for me, are a great expression of what reform should be. They go beyond the use of confidence intervals, or effect sizes, that seems to be so much of the reform debate. It's that broad approach that's almost totally missing from the new *Manual* (personal communication, February, 2002).

The heart of the issue, for most critics, is not that the *Manual* did not ban NHST, or mandate effect sizes, or prescribe any other particular methods. The heart of the criticisms is that the decision not to provide explicit requirements seems to have also excluded presenting the reasons for, and implications of, the recommendations. Following through with examples, offering general advice on good practice, and

providing explanations and education, are all things the *Manual* could have conceivably done without taking an official position on NHST.

It is difficult to think of reasons why the decision not to mandate or prohibit anything would rule out these other possibilities for encouraging reform. These absences cannot be attributed to external editorial pressures. Janet Hyde, then chair of the APA Publications & Communications (P&C) Board responsible for the production of the APA *Publication Manual* explained: "They [the TFSI] gave us a manuscript for that section that really looked good. We really didn't change much of it" (personal communication, August, 2001).

The preface of the *Manual* refers to the APA Style Website where changes to the fifth edition would be listed and where updates to the *Manual* would be posted (APA, 2001). In 2002, I published a version of this chapter and called for a substantial modification of statistics recommendations and examples to be prepared and posted on this website urgently. There have been no additions to the website since that time.

5.3 Downplaying Reform in the Promotion of the Fifth Edition

Criticisms of the *Manual* go further than comments on internal inconsistencies and missing material. The marketing of the *Manual* has also been a target because new recommendations related to statistical reform have been downplayed in publicity information for researchers and students. There has also been a general failure to promote the new guidelines to journal editors, authors, and students. Because a few new paragraphs in a very large new *Manual* might be easy for editors, authors, and students to overlook, the way that the changes in the *Manual* were highlighted in promotion was a critical ingredient in the *Manual* achieving impact. In the next section I outline how statistical reform was downplayed in promotion material.

5.3.1 2001 Promotion Examples

One of the most compelling examples of downplaying reform in promotion comes from the APA Style website, which summarises changes to "1.10 Results" of the fifth edition as follows:

> **Statistical presentation**: The description notes that the field of psychology is not of one mind about the reporting of "null hypothesis significance testing."
>
> **Informationally adequate statistics**: This section has been expanded to provide more guidance on reporting sufficient descriptive statistics.
>
> **Statistical significance**: In general and where appropriate, the exact probability, *p* value, should be reported.
>
> **Effect size and strength of relationship**: This section includes new information on multiple degree-of-freedom effect indicators and effect indicators[24]. (http://www.apastyle.org/chapter1.html, 02/15/02, bold in original)

Note that information that follows the Statistical Significance heading is detailed and prescriptive (i.e., "should be reported") whereas the information on effect sizes is descriptive only and does not relay specific changes to that section. CIs are distinctly absent from the list of changes—despite being recommended for the first time and described as "the best reporting strategy" (p. 22).

The back cover of the *Manual* suffers from similar problems. It contains two lists: The first is of “revised and updated items” and the second is titled “writers, scholars and professionals will also find”. The first list makes no reference at all to changes to statistics recommendations. The second refers only broadly to: "New guidelines on how to choose text, tables, or figures to present data" (APA, 2001).

Finally, the 2001 APA Convention in San Francisco hosted a workshop called "New Edition of the APA *Publication Manual*" (Knapp & Jackson, 2001). I attended this session of the APA Convention and collected all the available promotional material. This session covered general changes in the *Manual*—it was not restricted to changes in statistics recommendations. However, when changes to statistics were discussed, the discussion focused on the new recommendation to report exact *p* values to several decimal places. Other major changes, such as reporting CIs, were not raised at all.

---

[24] This may read as a misquotation, but it is not.

5.3.2 Journal Editors

There is consensus that journal editors need to be introduced to the new recommendations and that procedures need to be put in place to ensure they remain updated on further statistical developments. However, there seem to have been serious obstacles to this, one of which Janet Hyde outlined:

> The trouble is that editors tend to be very senior people in the field and they had their statistical socialisation a long time ago, so in a way they are the hardest ones to change. And yet they are the ones that have to lead on this and educate others... You've got these old, senior people who are the editors and some of them have, but many have not, kept up with new statistical developments. We have to figure out a way to educate them and persuade them that it is the right thing to do (personal communication, August, 2001).

Patrick Shrout (1997) was pessimistic about editors following through on the recommendations in the *Manual*:

> I honestly don't think so. Because I think the editors don't read that carefully. I think the people who read that [i.e., the *Publication Manual*] carefully are the less advanced people. So new students will read it but the power is being held by the people that have been schooled in the conventional way (Patrick Shrout, personal communication, September, 2001).

Geoff Loftus was also pessimistic about the impact of the *Manual* and about editors spontaneously engaging with the new recommendations:

> I think the *Manual* is not going to have as much effect with this [influencing statistical practice] as the individual editors. And the individual editors are not chosen on the basis of their statistical philosophy (Geoff Loftus, personal communication, August, 2001).

Joe Rossi is advocate of statistical power and meta-analysis (e.g., Rossi, 1990, 1997). As a reviewer, he has been requesting effect sizes from authors for many years. In many cases he is the only one to make the request: "I can tell from the other reviewers' comments that usually I'm the only one making that suggestion" (Joe Rossi, personal communication, September, 2001). Changes in the *Manual* should have made

such requests easier or at least more common. Yet, for him, there is little evidence this has occurred.

Richard Harris expected *some* recommendations to have an impact, but was far from reassured that the impact of effect size recommendations would be noticeable:

> My impression from what I've seen so far is that the *Publication Manual* takes a reasonable position somewhere between being rigidly doctrinaire and couching the emphasis on effect sizes so mildly that it has no real impact on practice (personal communication, January, 2002).

Frank Schmidt recalled several requests from reviewers or editors to cut effect size measures and CIs from his manuscripts. This happens to researchers, he claims, "more often than you think" (Frank Schmidt, personal communication, August, 2001). Schmidt suggested that the recommendations in the new *Manual* may be a useful defence for authors attempting to report data analyses they believe are important. However, he too is pessimistic about editors using them to request effect sizes and CIs from authors (Frank Schmidt, personal communication, August, 2001).

Others are substantially more optimistic about the impact the new recommendations will have on practice. Philip Kendall agreed that editors need to engage with the *Manual's* new recommendations. During his term as editor *JCCP* he was, as discussed in Chapter Four, one of very few editors to encourage the reporting effect sizes and measures of clinical significance, in line with changes he saw in the discipline (i.e., the TFSI and revisions of the *Manual*). Kendall argued that the institutional changes, as well as his policy, were having an impact on reporting in *JCCP*. However, the results reported in the previous chapter suggest this change may, at best, be very slow.

> The fact that the *Publication Manual* states the need to report effect sizes is a definite step in the right direction, but it will take co-operation from the editors to oversee that the movement actually takes place. I have seen an increase in both (a) papers submitted with effect sizes and (b) revisions, where effect sizes was requested, including effect sizes before the paper appears in print. (Philip Kendall, personal correspondence, February, 2002).

Roger Kirk was hopeful, but disappointed:

> I think we will see more effect sizes—no question about that. But that is not enough. So I'm very disappointed. I had such high hopes and they're

> just dashed. I read it and I thought 'there must be something more' (personal communication, August, 2001).

Further reason to be pessimistic about the influence of the fifth edition comes from Finch, Thompson & Cumming (2002) survey of the impact of APA *Manuals* (first to the fourth editions) on statistical reporting practice. They concluded:

> APA Manuals have not generally been responsive to authoritative calls for change and, although influential, have in some important cases not proved effective in shaping practices for reporting statistical analyses.

5.4 Other Criticisms

What I have provided here is not an exhaustive list of criticisms of the statistics recommendations in the *Manual*. A further criticism, for example, relates to the advocacy of standardised effect sizes. The 1999 TFSI report and the *Manual* present these measures as unproblematic, with no indication of the associated controversy. Patrick Shrout urged caution regarding the use of some of the effect sizes recommended in the *Manual*:

> I don't recommend R^2 that often either. Talking about the proportion of the variance that is accounted for is [problematic], at least in psychology [where it] can be much affected by how the sample is constructed. A lot of the time, people select samples that are not typically representative of some population... and if the treatment has an effect then the total variance is inflated. It is inflated as a function of the total number of people in one group versus the other. So it is not really coherent. You can use η^2, but I'm not sure what those numbers really mean and whether they can be manipulated by the design. If you have an effect, you can make the numbers seem large or small (Patrick Shrout, personal communication, September, 2001).

Geoff Loftus was also reluctant to endorse standardised effect size measures:

> ... [standardised] effect sizes always left me a little luke-warm, at best. Because they are... not connected to anything in the real world. So you have an effect and the size is defined by the standard deviation of the effect. You can sort of see why people would do that, but it doesn't seem

> to lend itself to cumulative progress... It has its place in the world, but it is kind of a minor place (personal communication, August, 2001).

Richard Harris was similarly unconvinced of the meaningfulness of standardised effect sizes:

> My particular "shtick"... is that effect sizes should, whenever possible, be couched in the original, "raw" units of measurement (or simple transformations thereof) rather than automatically being couched as a normalized *d* or R^2 measure. For instance, if you're doing a weight-loss study, the obvious effect-size measure is mean number of pounds lost. (personal correspondence, January, 2002).

In medical literature, standardised effect size measures have been severely criticised (Greenland, 1998). Ken Rothman, past assistant editor of *American Journal of Public Health* and past editor of *Epidemiology*, agreed that within epidemiology and medicine more generally, standardised effect sizes are widely considered not only meaningless, but also invalid (Ken Rothman, personal communication, August, 2001). There is also some coverage in the psychological literature (e.g., Richards, 1982). I will discuss these particular criticisms of standardised effect sizes again in Chapter Seven. However, it is worth noting here that this controversy goes unacknowledged in both APA *Publication Manual* and the TFSI report.

5.5 Summary and Conclusion

The statistics recommendations in the fifth edition of the APA *Publication Manual* now include strong recommendations to report confidence intervals and effect sizes. In addition, the controversy over NHST has been acknowledged (p. 21). There seems to be consensus that these changes are a genuine sign of progress, or at the very least, a step in the right direction.

However, the recommendations in the new *Manual* have disappointed many advocates of statistical reform. First, the examples in the *Manual* do not correspond with its own recommendations or with recommendations in the TFSI report; they continue overemphasis NHST at the expense of reform-related recommendations, such as statistical power, CIs, and effect sizes.

Second, there are inconsistencies in recommendations for and examples of NHST reporting itself. The examples presented in the *Manual* are the models of

reporting behaviour that researchers replicate in journals. Consequently, the new *Manual* does little to encourage change. Most agree that the overwhelming message is, as Finch, Thomason and Cumming (2002) put it, "NHST business as usual" (p. 841).

Other criticisms go beyond pointing out inconsistencies or absences in the *Manual's* examples or recommendations. Some critics were disappointed that the *Manual* did not make a pro-reform statement of principle. Others were disappointed that it did not take more of an educational role, by providing more complete explanations of *why* reporting effect sizes and CIs are important and *how* a failure to do so might damage the progress of psychology and related disciplines.

Given these failings, and that statistical reform was downplayed in APA material publicising changes, it is not at all clear that the *Manual* will affect statistical practices. The need for journal editors to engage with the recommendations and institute reformed policies within their own journals was raised frequently in the interviews I conducted. Yet, there has been an apparent lack of structure in place to ensure journal editors are adequately informed about the importance of newly recommended practices.

In its fifth edition, the APA *Publication Manual* has attempted to respond to calls for change. It has adopted some specific recommendations to report statistics that reformers have been advocating for many years. To this extent, reformers see some long-awaited progress being achieved. Yet, the failure to follow through on TFSI principles, and on the recommendations on its own pages, may itself serve only to undermine statistical reform.

6
STATISTICAL REFORM IN MEDICINE

After 17 years of interacting with physicians, I have come to realize that many of them are adherents of a religion they call Statistics… To the physician who practices this religion, Statistics refers to the seeking out and interpretation of p values. Like any good religion, it involves vague mysteries capable of contradictory and irrational interpretation (Salsburg, 1985, p.220).

In the mid 1980s there was a dramatic shift in the way statistical data were reported in medical journals. CIs, previously rarely reported, became routine. This shift occurred, in large part, through the efforts of journal editors. This chapter provides a chronicle of reform events in medicine, as Chapter Four did for psychology. Of course, medicine still has a number of statistical reform challenges to face. The discipline is far from a paradigm of perfect practice, as I explain later. But on a particular set of criteria—de-emphasizing statistical significance in favour of estimation in statistical reporting in journals—they are considerably ahead of psychology and can perhaps offer lessons.

6.1 Early Criticisms of NHST in Medicine

Early criticisms of NHST, and particularly its use in clinical trials, appeared in the medical literature in the mid 1960s (Cutler, Greenhouse, Cornfield & Schneiderman, 1966). Researchers began to worry that the technique was being too heavily relied upon. They were also concerned about professional boundaries, as clinicians became concerned that statisticians would claim authority over the conclusions of clinical trials (Marks, 1997).

In the 1970s criticisms of NHST became increasingly common (Band & Boen, 1972; Shulman, Kupst & Suran, 1976). Some critics of this era began advocating CI reporting instead of *p* values (Green, 1972; Wulff, 1973). To facilitate rapid reform, others provided guides to calculating CIs for relevant bio-medical effect sizes, such as odds ratios and relative risk values (Rothman, 1975, 1978a).

It was also around this time that statistical failings were first presented as an *ethical* concern. The pivotal ethical concern was underpowered clinical trials. May (1975) explained: "…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" (p. 23). Similarly, Newell (1978) pointed out that more clinicians ought to be aware of the ethical considerations related to statistical power: "Not every clinician—or even his ethical committee—is acutely attuned to the details of statistical Type II errors" (p. 534). Altman (1982a) went further, explaining that ethical problems are raised by both overpowered *and* underpowered studies:

> 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 (p. 6).

The problems Altman raised were indeed happening at the time. As we saw in Chapter Two, Freiman, Chalmers, Smith and Kuebler (1978) found *none* of 172 clinical trials published in the *New England Journal of Medicine (NEJM)* mentioned statistical power, type II errors or sample size calculations. Altman (1982b) specified three ethical consequences resulting from neglect of statistical power issues.

> (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 (p.1).

Certainly in *The Case of Intravenous Streptokinase for Acute Myocardial Infarction* (Chapter Three) all three of the above implications were played out in the testing of at least 30,000 unnecessary subjects, 15 years of virtually wasted research and the publication of confusing and apparently conflicting results. When we consider that medical research is often dealing with 'life and death' issues, the emphasis on ethical considerations of these statistical issues is not surprising.

In medicine, unlike psychology, these essentially statistical problems were immediately seen as having far broader implications. They were not 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. They were serious concerns—ethical concerns—for every researcher, statistician or not.

How the framing of these problems as an ethical concern eventually lead to advocacy of reporting CIs (rather than statistical power calculations, or any other NHST alternative for that matter) is not clear. (Current practice in medicine is that *a priori* power calculations are often used at the planning stage, especially in funding applications, and CIs are used for reporting of research findings. The latter are much more visible, but the former is admittedly widespread.) Douglas Altman, a reform leader in medicine (already quoted here) began advocating reports of statistical power (e.g., 1980) but by 1982 (and in subsequent articles and books, e.g., Gardner & Altman, 1989) was promoting increased reporting of CIs instead. CIs were already being heavily promoted by Ken Rothman and others, as we shall see below, by the mid to late 1970s.

Another major development of the late 1970s was the emerging tradition of quality control surveys of statistical reporting in medical journals. Larger journals, such as *NEJM* and *BMJ*, initiated their own investigations but other independent surveys also appeared in the literature. These studies often resulted in changes in journal policy, as the following sections demonstrate.

6.2 Journal Editorial Policies

New England Journal of Medicine

In the late 1970s, the prestigious *New England Journal of Medicine (NEJM)* instigated a review of its own statistical reporting practices. John Bailar and Fredrick Mosteller, in consultation with the *NEJM* editorial board, established a Study Group for Statistical Methods in the Biomedical Sciences. The Bailar-Mosteller group, as they came to be known, was largely made up of colleagues from the Harvard School of Medicine. The review generated so much interest that the project extended far beyond simply cataloguing reporting practices in journal, as was originally intended. In the end, the group published more than 30 articles and an edited book that reprinted some earlier articles, *Medical Uses of Statistics* (Bailar & Mosteller, 1986).

As part of the review, Emerson and Coditz (1983) surveyed *NEJM* and found that 44% of articles with *t* tests reported *p* values and 27% of articles with contingency tables reported *p* values (they did not report an overall percentage of articles with *p*

values). The frequency of *p* value reporting, despite being considerably lower than that in psychology journals, was interpreted as alarmingly high. Given the reporting frequency, and the controversy already surrounding their use, some members of the Bailar-Mosteller group wrote a paper devoted to clarifying their use (Ware, Mosteller & Ingelfinger, 1986). Ware et al. issued warnings about common misconceptions associated with *p* values and recommended that researchers instead report CIs, drawing particular attention to the fact that "the confidence interval gives more information than the *P*-value…and the width of the confidence interval gives an indication of the informativeness of the study" (p.156).

The group had the support of Arnold Relman, the then editor of *NEJM*. Relman (1986) wrote the preface to the edited book and endorsed the aim of the Bailar-Mosteller, which he described as "to tell us whether the methods [being used] were appropriately applied and how their use might be improved, and…to do so in simple language that would be understood even by readers who had no education in biostatistics" (p. xi).

There were other editorial initiatives at *NEJM* around the time the Bailar and Mosteller project was established, not the least of which involved Ken Rothman, who I discuss further in a later section. Rothman was on the *NEJM* editorial board in the late 1970s and in 1978 he had published the editorial "A show of confidence", advocating that CIs replace NHST. The same year, then Deputy Editor, Drummond Rennie published "Vive la difference ($p<0.05$)", also critical of the current dichotomous decision making practice based on NHST. The close timing of these events is important—offering the chance for each to reinforce the others.

Then, as now, *NEJM* was a highly respected medical journal. It currently has the highest impact factor rating of any medical journal: 38.570. Having such a prestigious journal take the bold first step of investigating statistical reporting and advocating new practices did not necessarily guarantee successful reform in medicine, but it certainly helped promote the cause.

British Medical Journal

Sentiments similar to those above have been expressed about reforms in the *British Medical Journal (BMJ)*: "The widespread readership of the *British Medical Journal* has been cited as one reason as to why it has been Gardner and Altman who have taken the credit for the better reporting of statistics today" (Rigby, 1999, p.714).

Whilst internal reviews were taking place at *NEJM*, the *BMJ* was under similar scrutiny. In the late 1970s, *BMJ* published an article reviewing the misuse of statistical methods in the journal (Gore, Jones & Rytter, 1977). The survey found several deficiencies, including failure to: a) indicate what hypotheses were being tested; b) provide degrees of freedom or sample size and c) report measures of central tendency, such as means or medians or measures of spread, such as standard deviation. In addition, there were routine failures to meet the assumptions of Student's *t* tests and chi-square tests. The same year, *BMJ* also published a theoretical critique of NHST (Petro & Doll, 1977).

In the early 1980s the controversy over use of *p* values in *BMJ* grew. *Statistics in Practice*, a series of *BMJ* articles by Douglas Altman and Shelia Gore, was published by *BMJ* books in 1982. This was the *BMJ* equivalent of the slightly later *NEJM*'s *Medical Uses of Statistics* (Bailar & Mosteller, 1986). The articles reprinted in the *BMJ* collection dealt with a wide range of statistical issues, including deficiencies of dichotomous decision making inherent in NHST, neglect of statistical power and the advantages of CIs. Altman's section, discussed earlier in this chapter, focused on the ethical implications of statistical deficiencies. Gore's section dealt with the presentation and interpretation of results. For example, in "Assessing methods: Art of significance testing", Gore warned researchers not to over-interpret NHST: "clinical decisions should not be made automatically on the basis of a single 'statistically significant' finding" (1982, p.601).

Altman (1982c) strongly emphasised the responsibility journal editors and manuscript reviewers should take in improving statistical practices in journals. He argued that "all papers using any statistical procedure should be refereed by a statistician" (p.22), and further, that they should be sent back to the statistical referee after any changes for re-checking. Altman's proposal was thorough. Journals should be obliged to state clearly their policy; they should offer statistical guidelines to contributors, and they should employ editorial staff with statistical competency. Finally, authors should be obliged to make full data sets available and supply referees with additional information regarding statistical methods used, including copies of related papers.

At the same time, related empirical questions about review process were being addressed. As part of a then newly emerging tradition of surveying statistical practice Gardner, Altman, Jones and Machin (1983) asked "Is the statistical assessment of

papers submitted to the *British Medical Journal* effective?" Their answer was a resounding 'no'.

The *BMJ* publications established misuse and misinterpretation of NHST as not only an *ethical* concern, but also an *editorial* concern. In response to growing criticisms, and the undeniable deficiencies in statistical reporting in the journal, *BMJ* adopted the following policy on CI reporting: "…from 1 July authors of papers submitted to the *BMJ* will be expected to calculate confidence intervals whenever the data warrant this approach" (Langman, 1986, p. 716). In 1989, Langman's policy was reinforced by Gardner and Altman's text *Statistics with Confidence*, also published by *BMJ* books. In their introduction Gardner and Altman noted the spread of this type of statistical reform in other medical 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 (1989, p.4).

(I discuss the International Committee of Medical Journal Editors next in this chapter.) Gardner and Altman not only explained the benefits of CIs, they also gave worked examples of how to calculate CIs for regression, correlation, relative risks, odds ratios, survival time analyses and some non-parametric analyses. Their text included software for carrying out such calculations (Confidence Interval Analysis, CIA). They were clear about the need for such guidelines:

> One of the blocks to implementing this policy [on CIs at *BMJ*] has been that the methods needed to calculate confidence intervals are not readily available in most statistical textbooks. The chapters that follow present appropriate techniques for most common situations. Further articles in the *American Journal of Public Health* and the *Annals of Internal Medicine* have debated the uses of confidence intervals and hypothesis tests and have discussed the interpretation of confidence intervals (1989, p.4).

The obstacle Altman and Gardner acknowledge here was an important one. And their solution to it, a timely publication of a statistical text, was, as I argue in Chapter Seven, a crucial development.

Seldrup (1997) surveyed reporting practices in *BMJ* before and after its 1986 editorial policy on reporting CIs. The proportion of articles with CIs rose as a consequence of the policy—from just 4% to 62%. The policy had clearly been influential, and the proportion of articles reporting CIs has continued to increase: In 2003, over 80% of empirical articles in *BMJ* reported CIs (Coulson, Fidler & Cumming, 2005).

The International Committee of Medical Journal Editors

Before moving on to discuss two other medical journals, I need to introduce the International Committee of Medical Journal Editors (ICMJE). This committee had met annually since the late 1970s, when they were known as the 'Vancouver group'. It was the original group's charter to devise some uniform editorial requirements for medical journals. In 1988 they became directly involved in statistical reform in medicine when they revised their "Uniform Requirements for Manuscripts Submitted to Biomedical Journals" to include the following statement regarding statistical inference:

> 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 (p.260).

The "Uniform Requirements" were published in the *Annals of Internal Medicine* (ICMJE, 1988a) and the *BMJ* (ICMJE, 1988b). Over 300 medical and biomedical journals notified ICMJE of their willingness to comply with the manuscript guidelines.

Although the ICMJE guidelines were relatively brief on issues of statistical inference—the above quotation is not far from the whole piece on this subject—they were timely. In less than five years several leading journals (including *NEJM, BMJ, Lancet, AJPH* and others) had all adopted reform policies and they had quickly received the institutional support of the ICJME. The importance of the timing of this institutional support is discussed again in Chapter Seven. Partly making up for the actual guideline's brevity, was the accompanying the *Annals of Internal Medicine* publication by Bailar and Mosteller (1986): "Guidelines for statistical reporting in

articles for medical journals: Amplifications and explanations". As the title suggests, this article expanded on the ICMJE guidelines, and offered further advice and guidance on implementing new recommendations.

6.2.1 A Case Study in Editorial Policy: Ken Rothman's reforms at the American Journal of Public Health and Epidemiology

By the time he became assistant editor of the *American Journal of Public Health (AJPH)* in 1983, Ken Rothman had been advocating statistical reform for almost a decade. As I mentioned earlier, he had previously been on the board of *NEJM* where he had published an editorial recommending CIs replace *p* values (Rothman, 1978); earlier still he wrote computational articles for using CIs (e.g., Rothman, 1975). After completing his term at *AJPH* he wrote *Modern Epidemiology* (Rothman, 1988) which quickly became an influential and widely used advanced statistical text.

Rothman's contributions to reform in epidemiology, and medicine more generally, are widely acknowledged. Of Rothman's policy at the journal *Epidemiology* Altman (2000a) wrote: "I am unaware of any other medical journal which has taken such a strong stance against P values." (p. 9). Charles Poole, also a prolific critic of NHST and advocate of CIs, explained how particularly influential Rothman was in statistical reform of epidemiology:

> I believe the change in statistical reporting practices in US epidemiology occurred because of one person: Ken Rothman. He was president, incoming and ongoing, of almost every one of our societies. He wrote the first sophisticated methodologically orientated textbook [*Modern Epidemiology*], he's a wonderful speaker, extremely charismatic. I think personality had a lot to do with this reform… personality and force of persuasion (Charles Poole, personal communication, September 2001).

At *AJPH* Rothman took the most radical stance that had yet been taken in statistical reform. In his revise and submit letters to would-be authors in *AJPH* 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 by Fleiss, cited by Shrout 1997).

Not surprisingly, this created controversy. Fleiss (1986) wrote to *AJPH*: "An insidious message is being sent to researchers in epidemiology that tests of significance are invalid and have no place in their research" (p.559). Others agreed in principle with a shift from NHST to CIs (Rothman's preferred alternative), but were unhappy with the process used to implement the change. To re-quote Patrick Shrout, then of the Columbia School of Public Health: "We were outraged that this happened overnight…These poor epidemiologists who suddenly had the rules changed. Ironically, we were in sympathy with the goal, but we resented the heavy-handedness" (personal communication, August 2001).

In Rothman's defence, by the time he started with *AJPH*, there had been substantial criticism of NHST and discussion of CIs in mainstream medical literature (several examples are cited earlier in this chapter). Furthermore, Rothman himself had made his views on this topic public, long before his *AJPH* appointment. His own view of his editorial activities at *AJPH* is not one of 'heavy handedness':

> My revise-and-resubmit 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 correspondence, July 2002).

Rothman pointed out however that when he became the founding editor of *Epidemiology* his policy was both stricter and more explicit than any earlier policies. In an editorial for this journal he wrote:

> When writing for *Epidemiology*, you can enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals discourages them, and every worthwhile journal will accept papers that omit them entirely. 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)

In 2001 I undertook a joint research project (published as Fidler, Thomason, Cumming et al., 2004) to survey the effectiveness of Rothman's editorial policies, both at *AJPH* and later at *Epidemiology*. We surveyed *AJPH* articles published before, during and after Rothman's policy, before and after the ICMJE regulations, and before and after changes to the "Instructions to Authors". We coded 594 *AJPH* articles, published in selected years between 1982 and 2000 (see Table 6.1). From *Epidemiology* we coded 40 articles published in 1990, the year Rothman founded the journal, and 70 from 2000, his final year as editor. We coded articles with new data only; we did not code meta-analyses, methodological or theoretical articles. We recorded a practice (e.g., CI use) as present if an article contained at least one instance of the practice; we did not count any further instances.

Items Coded

NHST. We coded whether NHST was used and instances where the author did not clarify whether 'significant' meant 'important' or 'statistically significant'. If the author did not: (a) preface 'significant' with 'statistically', or (b) follow the statement of significance directly with a *p* value or test statistic, or (c) otherwise differentiate between statistical and substantive interpretations, then the practice was recorded as ambiguous. We coded whether the author reported the relevant test statistic (e.g., *t* or *F* value) for any significance test, as is needed for full reporting of NHST.

Statistical Power. If a power calculation was reported we coded 'explicit power'. Otherwise, we searched for any mention of the relationship between sample size, effect size and statistical significance (e.g., a reference to small sample size as perhaps explaining failure to find statistical significance). This was coded as 'implicit power'.

Confidence Intervals. We recorded whether CIs were presented in text, table or figure, and whether they were interpreted. Interpretation included any mention of interval bounds or width, any reference to interval overlap, or reference to the null value being inside or outside an interval.

Effect Sizes. We coded reports of any effect size—means, odds ratios (ORs), relative risk values, percentages, proportions, regression coefficients, correlation coefficients, standarised effect sizes (such as Cohen's *d*), other units-free measures such as η or η^2, ω or ω^2, and variance accounted for statistics, such as R^2.

Table 6.1.
Publication years chosen for coding *American Journal of Public Health* articles, number of articles coded, and reason for interest in those years.

Year	Number of articles coded	Reason for choosing year
1982	67	Pre-Rothman
1986	98	Expected maximum influence of Rothman whose term was 1984 to February 1987
1988	71	Immediately post-Rothman
1989	72	Post-Rothman
1990	72	Post-Rothman and post-ICMJE recommendations (published 1988 and referred to in *AJPH* "Instructions to Authors" in 1989)
1993	72	Editor change and ICMJE recommendations dropped from *AJPH* "Instructions to Authors" in 1991
1994	72	As for 1993
2000	70	Recent practices

Reliability of coding. A random selection of articles was independently recoded: 48 of 594 from *AJPH*, and 10 of 110 from *Epidemiology*. The accuracy of the original coding was 92%. Errors were almost exclusively missed reports (so frequencies reported here may be slight underestimates) and were distributed approximately evenly across all categories.

Results

NHST. Of the 594 *AJPH* articles, 273 (46%) reported NHST. In almost two thirds (64% of 273) 'significant' was used ambiguously. Relevant test statistics were reported in only 104 (38%) articles. Only 8 (3%) articles reported explicit statistical power, and an additional 42 (15%) implied power. Thus, an overwhelming 82% of NHST articles had neither an explicit nor implicit reference to statistical power, even though almost all reported at least one statistically non-significant result. In *Epidemiology*, only 4 of 110 articles reported NHST.

Confidence Intervals. Of the 594 *AJPH* articles, 322 (54%) reported CIs; of 110 *Epidemiology* articles, 95 (86%) had CIs. The overwhelming majority of *AJPH* articles (268 of 322, 83%) reported CIs in tables (often very large ones); only 15 (5%) displayed error bars in figures. In *Epidemiology*, the corresponding frequencies were 81 (85% of 95) and 6 (6%). Table 6.2 shows that fewer than 12% of *AJPH* articles with CIs interpreted them. Despite 86% of articles in *Epidemiology* reporting CIs, interpretation was in that journal (at least according to our criteria) was also rare. Nor did the situation improve over time. For example, only 1 (of 40) 1990 *Epidemiology* paper referred to CI width; in 2000, only 3 (of 70 papers) did this.

Table 6.2.
Type and frequency of confidence interval interpretation in *American Journal of Public Health and Epidemiology*.

CI Interpretation	*AJPH* percentage (number) out of 322	*Epidemiology* percentage (number) out of 95
Any mention of CI limits	1.2 (4)	1.1 (1)
Any mention of CI width	2.2 (7)	4.2 (4)
Any mention of CI overlap	1.9 (6)	0
Any reference to null value	6.2 (20)	3.2 (3)
Any CI interpretation*	11.2 (36*)	8.4 (8)

*One article had two interpretations

NHST Versus CIs (1982-2000). Figure 6.1 shows that sole reliance on *p* values dropped dramatically during Rothman's term at *AJPH*, from 63% in 1982, to 6% in 1986-9. CI reporting increased from 10% before Rothman, to 54% in 1986 towards the end of his editorial service. As shown in Figure 6.1, the changes in percentages mentioned here are large by comparison with the 95% CI widths. In *Epidemiology*, CIs were even more common: 94% of articles in 2000 reported them. By contrast, *p* values were rare; there were none in 2000.

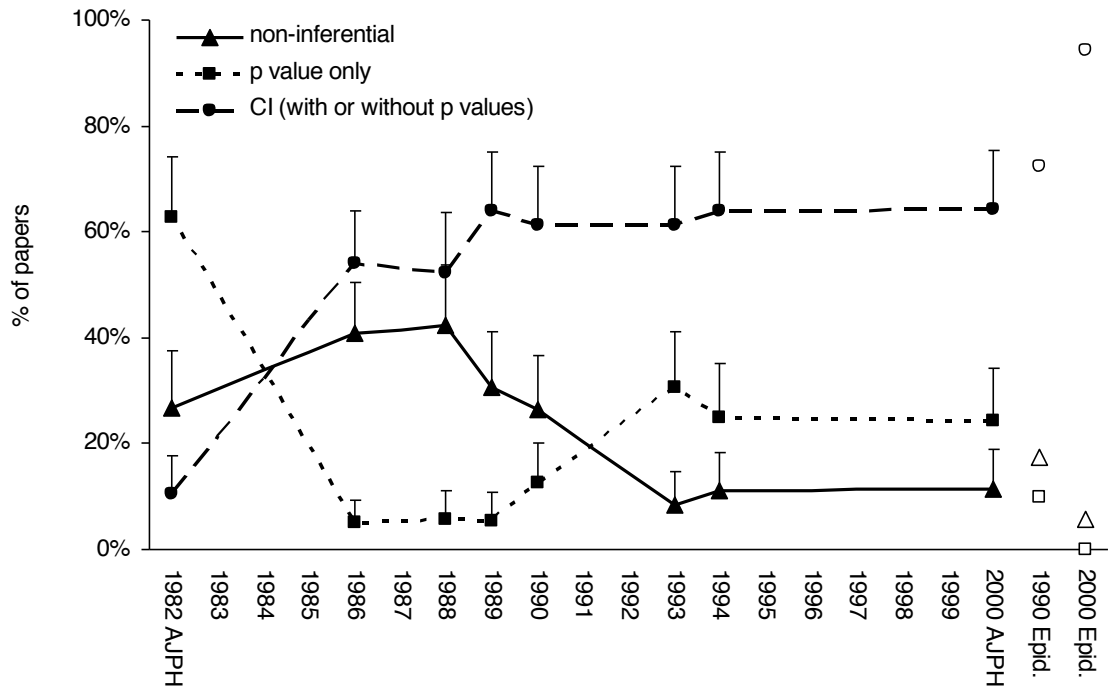


Figure 6.1. Percentage of *American Journal of Public Health* and *Epidemiology* articles reporting NHST, CIs or descriptive statistics only (non-inferential) between 1982 and 2000. Error bars are upper half 95% CIs.

Figure 6.1 also shows a concomitant increase from 1982 to 1986-8 in non-inferential articles in *AJPH*. These articles often reported effect sizes (e.g., means, percentages, ratios), but included neither NHST, CIs nor other inferential analysis. Some were legitimate descriptive studies of entire populations, and so did not require inference. Others, however, contained evidence that authors were doing *covert* significance testing. Occasionally there was explicit evidence for this, for example a footnote explaining that NHST, whilst not reported, had been conducted and readers were invited to contact authors for results. In other articles there was ambiguity. Although no NHST results were reported, discussions focused on 'significant differences'.

Figure 6.1 suggests the CI reporting in *AJPH* has been relatively stable since Rothman left. However, Figure 6.2 tells an interestingly different story. It shows that, while Rothman was at *AJPH*, CIs were commonly reported without *p* values. For some time after his departure in early 1987 this trend remained. By 1993, however, the number of articles reporting *p* values had increased dramatically. CIs continued to be reported, but from this point on were supplementing *p* values, rather than replacing

them. For example, in 1990, 42% of articles reported only CIs and a further 19% reported both CIs and *p* values. In 1993, these figures were virtuallyreversed: 13% reported only CIs and 48% reported both CIs and *p* values.

There was also an increase in sole reliance on *p* values. In 1990, less than 13% of articles relied only on *p* values; by 1993, this figure had more than doubled at just over 30%. This resurgence in *p* values followed the arrival of a new editor, and 1991 removal of the ICMJE's recommendations from *AJPH's* "Instructions to Authors".

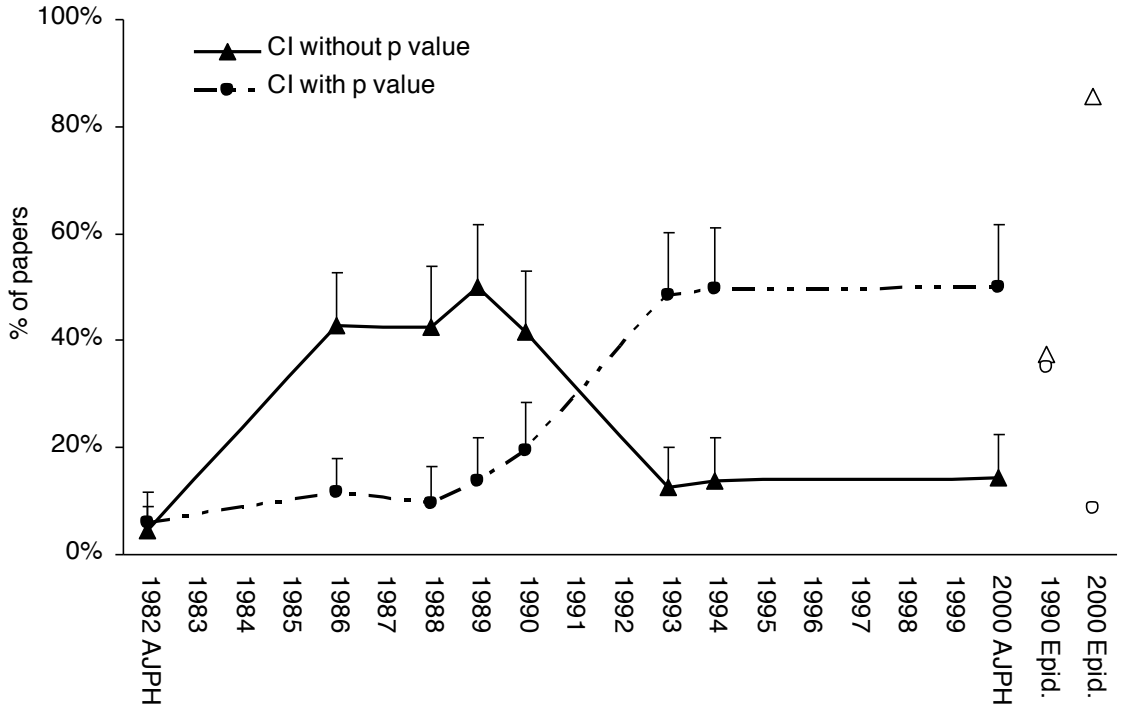


Figure 6.2. Percentage of *American Journal of Public Health* and *Epidemiology* articles using CI to replace NHST (CI without p value) or inconjunction with NHST (CI with p value). Error bars are upper half 95% CIs.

Effect Sizes. Figure 6.3 shows that effect sizes were very often reported as percentages and proportions (*AJPH* 95%*, Epidemiology* 84%). Means and mean differences were common in *Epidemiology* (92%) but were not as common in *AJPH* (38%). Odds ratios and relative risk values were reported in approximately half of the articles in both journals (*AJPH* 46%, *Epidemiology* 53%). Other units-free measures were reported only occasionally (e.g., in *AJPH* 13% of articles reported correlation coefficients, 9% reported R^2 values). There were no reports of effect sizes in standard

deviation units (e.g., Cohen's *d*) in either journal; rates were stable over the years surveyed.

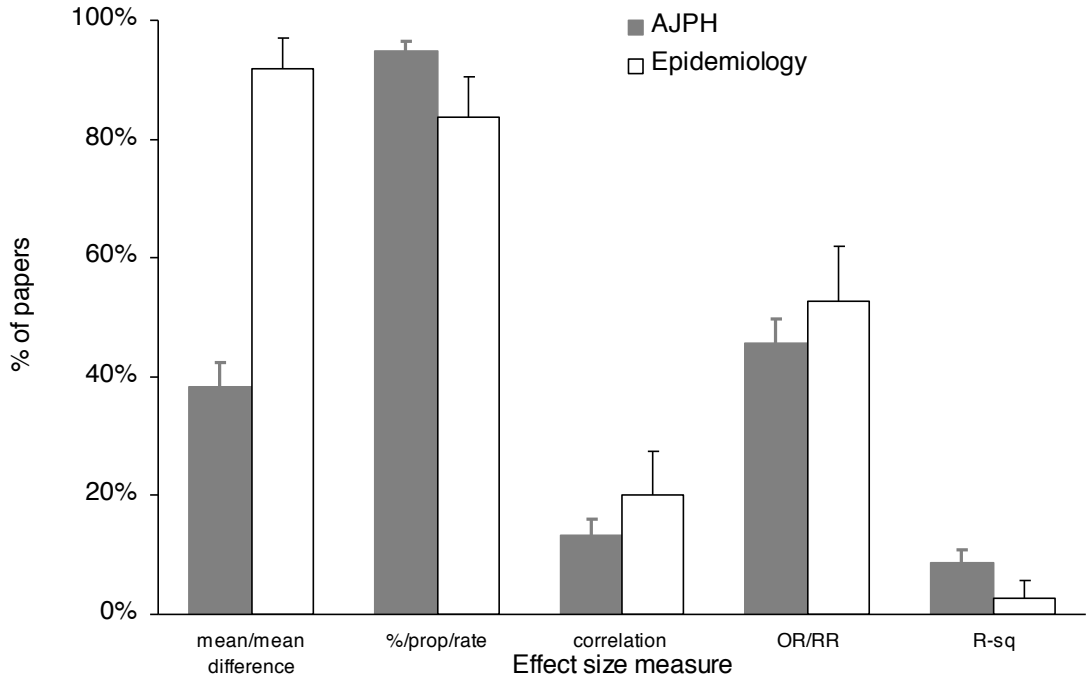


Figure 6.3. Percentage of *American Journal of Public Health* and *Epidemiology* articles reporting various types of effect size measures. Error bars are upper half 95% CIs.

Discussion of Survey Results

These data, as summarised by Figures 6.1 and 6.2, suggest that Rothman's efforts at *AJPH* were initially effective, leading to a remarkable drop in NHST use and increase in CI reporting. Several years after his departure CI reporting remained high—as it had become in many other medical journals—but *p* values had again become common in *AJPH*. He was more consistently successful at *Epidemiology*.

In both journals, however, when CIs were reported they were rarely used to interpret results or comment on precision. This rather ominous finding holds even for the most recent years we surveyed. In addition, in many *AJPH* articles in which NHST and *p* values were not explicitly reported, there was evidence, or at least clear hints, that interpretation was based on unreported NHST. Almost all articles reported some effect sizes, often percentages and ratios. Cohen's *d* and similar standardized effect size measures that are often needed in psychology for meta-analysis were not used. This is

perhaps not surprising given that units of measurement in medicine are more often universal (than, say, in psychology) and that, as mentioned earlier, there have been criticisms of standardised effect sizes in the medical literature (Greenland, Schlesselman & Criqui, 1986; Greenland, 1998).

6.3 The Food and Drug Administration, Pharmaceutical Companies and Funding Agencies

Salsburg (2001) noted that it was a rigid form of Neyman-Pearson hypothesis testing that made its way to the U.S. Food and Drug Administration (FDA). I have not been able to determine when statistical significance first became a requirement for new drug approval, despite several enquiries to the FDA's information department. Without that information it is difficult to find evidence for the hypothesis I propose. However, it is plausible that interest in statistical power developed concomitantly with the FDA's requirement. In an industry where a great deal of money depends on whether a drug produces a statistically significant effect (regardless of the size of the effect), the pressure on pharmaceutical companies to achieve statistically significant results would have no doubt raised the awareness of statistical power.

Another reforming force at work in medicine was the ethical framing of statistical power calculations in clinical trials involving human subjects. (I highlighted this point earlier.) Ethics committees and research funding agencies eventually moved to require statistical power calculations (or sample size calculations based on statistical power or precision) as part of their application process. To requote Maxwell (2004): "research in the health-related journals tends to be federally funded, and federal funding agencies may be likely to require evidence of sufficient statistical power before deciding to fund a proposal." (p. 148). This is perhaps not surprising, given the much higher average cost of medical research compared with psychological research. In the USA, the FDA's guidelines for conducting clinical trials now devote considerable attention to sample size calculations and CIs (FDA, Centre for Drug Evaluation and Research, *Guidance for Industry, E9 Statistical Principles for Clinical Trials*, http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF, last accessed 18/02/05).

6.4 A Way to Go?

'The Rothman'
It was 1997, at the Society of Epidemiological Research meeting: I went through the abstract book. There were 300 or so abstracts and was trying to find one that said anything at all about the width of a confidence interval. And I was prepared to be extremely charitable—if they had said 'these results were imprecise, but nevertheless suggest…'—it would have been enough. But I found none; zero! So, [in the hope of encouraging this kind of discussion] I set up an award. When I announced it, I called it 'The Rothman'. The next year I read all the abstracts and, again, zero abstracts said anything about confidence interval width. (Charles Poole, personal communication, September 2001)

Although CI reporting has been routine practice in most medical journals for almost 20 years, serious challenges remain. Despite widespread reporting of CIs in many journals the uptake of confidence intervals has not been equal throughout sub-fields of medicine (Altman, 2000a, 2000b). In a review of 1996 articles in the *American Journal of Physiology* only 1 out of 370 papers reported a CI (Curran-Everett, Taylor, & Kafadar, 1998). Similarly, only 2 of 112 articles in 1991-92 anaesthesia journals reported CIs (Mantha, Thisted, Foss, Ellis & Roizen, 1993). That these cases remain is a problem for reform, but perhaps not as serious a problem as the next issue.

In the case of *AJPH*, CI reporting increased dramatically but this did little to change the way researchers interpret and discuss their findings. Even in some articles that did not include any *p* values, discussions often still focused on 'statistical significance'. 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). As Poole explained of epidemiology:

> 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] wrote. But basically they are only reporting CIs because Ken was an authority figure and he and others encouraged them—well, his journal [*Epidemiology*] insisted on it. (personal communication, September 2001).

This raises important questions about what statistical reform is. In my opinion, change in the statistics reported is merely the first step. Substantial reform must also require changes in the way researchers approach analysis, and interpret and think about data. It requires cognitive change, and entails more than mechanically adding CIs to tables. Measured against this standard, medicine does indeed have a way to go. However, as I said at the start of this chapter, even it is far ahead of psychology.

In recent years, there have been fewer articles in medicine criticising NHST. This may indicate, as Altman (2000a) suggested, either that confidence intervals are fully integrated into statistics courses and routinely reported *or* that there is a misperception that "this particular battle has been won" (p.7). He concluded: "Probably all of these are true to some degree." (p. 7).

6.5 Summary

Though strict editorial policies and the timely rewriting of textbooks (and possibly pressure from regulatory and funding agencies), statistical practices in medicine were vastly changed by the mid 1980s. CIs were routine reporting practice in most journals. As I have mentioned, in a recent survey of 10 leading medical journals CIs were found in around 85% of articles (Coulson, Fidler & Cumming, 2005). In psychology, on the other hand, NHST continues to fill the journals, despite almost half a century of critics and, as in medicine, editorial and institutional interventions. CIs appear in only 10% of articles, in roughly similar types of articles (Coulson, Fidler & Cumming, 2005).

The answer to why these disciplines have progressed differently is not straightforward. There are various sociological factors that go some way to explaining why medicine had a statistical reform and psychology didn't. For example, in medicine, reform was led largely by journal editors, who collaborated regularly. Furthermore, institutional support from the ICMJE followed quickly the first individual changes in journal editorial policy. Similar support from the APA started over a decade later. Finally, ethical concerns related to statistical power have not become an issue to anywhere near the same extent in psychology.

7
WHY MEDICINE REPORTS CONFIDENCE INTERVALS AND PSYCHOLOGY DOESN'T

If one accepts the criticisms of significance tests… then it would appear that psychologists and other behavioural and social scientists, have, for the last 40 years or so, been almost wilfully stupid. What explanations can be offered for their failure to acknowledge, at a much earlier date, the cogency of these arguments? (Oakes, 1986, p.68)

In this chapter I attempt to answer Oakes' question. It is a question has been paid curiously little, sustained attention in the reform literature. Occasionally, some authors make an aside attempt to make sense of the lack of reform in psychology. These explanations almost always fall back on 'the intuitive nature of statistical fallacies'—that is fallacies about probability and statistics that lead researchers to misinterpret NHST, such as those discussed in Chapter Two.

In this chapter I first outline the inadequacies of fallacies and misconceptions to alone account for the lack of reform in psychology. Following this I outline Oakes' own answer to the above question, which includes but go beyond the fallacy explanation. I also look at other explanations offered by, for example, Thompson (1999) and John (1992). Individually, each of these explanations does help make sense of psychology's position. Yet, to varying degrees they are all subject to the same objection, that is, they cannot account for why medicine changed reporting practices and psychology did not.

Later in this chapter I provide a list of sociological factors that go some way to explaining differences between the disciplines. These factors help explain how medicine changed its reporting practices—for example, they enforced strict editorial policies and re-rewrote textbooks. However, what these factors mostly demonstrate is the difference between the disciplines in *receptiveness* to reform, not necessarily *motivation* for reform. Understanding medicine's motivation for reform, I believe, returns us to the way problems of inference were, from the beginning in medicine, framed as ethical problems, not merely technical statistical concerns. It no doubt also involves a better understanding of the influence of pharmaceutical companies, regulatory authorities (such as the FDA) and funding agencies. Such things are beyond the scope of this thesis.

The final explanation of lack of reform in psychology given in this chapter is more philosophical than sociological. It is about the difficulty of a move to estimation and precision in a discipline where there are so rarely natural or universal units to estimate in.

In summary, this chapter has the following structure. First, I critique the typical 'fallacy' explanation for why psychology hasn't reformed statistical practices. Second, I summarise and critique Oakes' and others attempt to answer this question. Third, I look at sociological differences between medicine and psychology in terms of their receptiveness to reform. Finally, I look at some conceptual-cum-statistical difficulties of instituting a reform based on estimation in disciplines like psychology.

7.1 Inferential Fallacies and Institutional Inertia

As I have explained, fallacies and misconceptions regularly lead researchers to misinterpret *p* values, in particular to assume that a *p* value provides a lot more information than it actually does. Because researchers believe they know things they don't, they then often fail to calculate and report other important statistics. The inverse probability fallacy provides an example. If a researcher commits the inverse fallacy, they interpret their *p* value as the probability of the null hypothesis, given the data (rather than the correct probability of the data given the null). This then becomes abbreviated to simply the probability of the null hypothesis, and therefore $1 - p$ becomes the probability of the alternative hypothesis. To the researcher committing the fallacy, *p* represents not only the probability of their statistical hypothesis being true, but also the probability of the substantive theory behind the hypothesis being true. If one knows the probability of their substantive theory being true, what interest could other statistics possibly hold? As Oakes said:

> …psychologists typically give a Bayesian interpretation to analyses that are strictly frequentist. The point is that 'power' is not a concept of any central interest to a Bayesian—after all why worry about the probability of obtaining data that will lead to the rejection of the null hypothesis if it is false when your analysis gives you the actual probability of the null being false? In short, I am suggesting that power is neglected by psychologists because, given their typical mistaken misunderstanding of statistical significance, it is an unnecessary concept (1986, p.83).

The empirical evidence that researchers do commit these inferential fallacies is uncontroversial (Haller & Krauss, 2002; Oakes, 1986; Tversky & Kahneman, 1971). So too is the evidence that misinterpretations of *p* values are frequently published in psychological literature (Fidler, Cumming, Thomason et al., 2005; Finch, Cumming & Thomason, 2001; Vaache-Haase et al., 2000).

Schmidt and Hunter (1997) have also invoked psychological explanations for the persistence of NHST. They collected around 80 objections to the discontinuation of NHST in psychology, that is, false beliefs that work to maintain current practice: "Each of these objections is intuitively appealing and plausible but is easily shown to be socially and intellectually bankrupt"(p. 37). The 8 most common objections were:

> (a) Without significance tests we would not know whether a finding is real or just due to chance;
> (b) hypothesis testing would not be possible without significance tests;
> (c) the problem is not significance tests but failure to develop a tradition of replicating studies;
> (d) when studies have a large number of relationships, we need significance tests to identify those that are real and not just due to chance;
> (e) confidence intervals are themselves significance tests;
> (f) significance testing ensures objectivity in the interpretation of research data;
> (g) it is the misuse, not the use, of significance tests that is the problem; and
> (h) it is futile to try to reform data analysis methods (p. 37).

Inferential fallacies, misconceptions about *p* values and misguided objections have no doubt played a significant role in maintaining the statistical status quo in psychology. We know that psychologists hold such misconceptions, and in the ways suggested here, they work to maintain researchers' investment in current practice.

Yet, there is a problem with this explanation as a full account of why reform has not occurred in psychology. Medical researchers also hold such misconceptions and medicine did reform, or at least change reporting practice. There is certainly no *a priori* reason to suspect medical researchers would be any less susceptible to the common misconceptions and fallacies, and results from empirical studies quickly put to rest any lingering doubts about this. For example, studies conducted before or around the time of major reporting changes in medicine demonstrate, not only misconceptions in the

literature (Ambroz, Chalmers & Smith, 1978; Feinstein, 1974; Freiman et al., 1978; Gore, Jones & Rytter, 1977; Pocock, Hughes & Lee, 1987; Rigby, 1998; Schor & Karten, 1966) but also direct evidence of researchers' misunderstanding (Borak & Veilleux, 1982; Wulff, Andersen, Brandenhoff & Guttler, 1987). The key question, therefore, becomes not simply "Why hasn't psychology improved its statistical practices?" but rather "Why hasn't psychology managed it when medicine has?"

7.2 Oakes' Explanations for the Longevity of NHST in Psychology

Oakes offered the following three explanations for the persistence of NHST in psychology: (1) inertia and submission to statistical authority, (2) the weakness of the proposed alternative and (3) the prevailing philosophical climate (1986, p.68-72). Below I outline each of these in turn.

First, of inertia Oakes wrote: "Social scientists, like human beings, are creatures of habit" (p.68). In other words, we should find nothing strange or unusual about resistance to change—it is a human condition. Science is run by humans and therefore no exception. Paul Meehl also attributed the persistence of NHST to 'plain psychic inertia':

> …plain psychic inertia is a powerful factor in science, as it is in other areas of life—don't underestimate it. When the issue is method, rather than substance, it makes it worse. If one has been thinking in a certain way since he was a senior in college 'the way you test theories in psychology is refute H null', there is a certain intellectual violence involved in telling a person, well, not that they've been a crook, but that they've been deceiving themselves (personal communication, August 2002).

There can be little doubt that Oakes and Meehl are right that inertia has been a powerful factor in the longevity of these statistical practices in psychology. Without downplaying the importance of this factor, it is important to ask whether researchers in psychology suffer more from 'plain psychic inertia' than researchers in medicine. Only extensive personality testing could tell for sure, although there seems no prime-face reason to believe psychologists would be more susceptible to such forces. How then was medicine able to overcome the inertia and psychology not? A complete answer to

this question must take account of (at least) the sociological factors that I list in this chapter.

Of submission to statistical authority Oakes speculated that "many researchers are in awe of their statistically minded colleagues and even more so of the thoroughgoing academic statistician" (p.69). He pointed out that such submissiveness is not particular to psychologists; that many statisticians also found it difficult to accept there could be mistakes in the work of the genius Fisher, or Neyman and Pearson. Again, I agree with Oakes that this factor has contributed to maintaining the situation in psychology, and again it seems unlikely then such a trait would not exist amongst medical researchers.

Oakes' second factor—the weakness of the proposed alternative—suggests that criticisms to NHST have been met with apathy because they are perceived to entail a commitment to the subjectivity probability of the Bayesians. Oakes pointed out however, that most criticisms of NHST can be made with indifference to Bayesian methods, that is, they can be made just as convincingly from within a frequenist framework.

Oakes is indeed right that most criticisms can be made from within a frequentist framework. In fact, the vast majority of criticisms in psychology and medicine *have* been made from within a frequentist framework (as Chapter Two demonstrates). Many researchers exposed to such criticisms are possibly not even aware that a Bayesian school of probability exists. I therefore remain sceptical of the impact of this particular factor. In making such a proposal Oakes has, I believe, overestimated at least the awareness of Bayesian methods, if not resistance to them.

Finally, Oakes suggested that the prevailing philosophical climate of Popperian falsification has kept NHST dominant for so long.

> If theories can only be falsified, and thereby lies knowledge, isn't that exactly what significance tests do? ... Hence I am suggesting that a contributing factor in the retention of significance tests is their superficial philosophical respectability (p.70-71).

NHST and falsificationism agree that theories ought to be tested, so there is a resemblance between Fisherian and Neyman-Pearsonian NHST and Popperian philosophy. However, the resemblance could hardly be more trivial. Granted, the language of NHST does conveniently map on to Popper's:

- Theories can only ever be shown to be false (we can *reject* the null);
- The survival of a test does not prove a theory (we never *accept* the null hypothesis);
- Survival merely means the theory is corroborated, and more so the more tests it survives (we continue to *fail to reject*).

Yet this 'fit' is at best superficial, at worst, an illusion. In more serious ways, NHST violates Popper's falsifiability requirements. Popper's (1959, 1962) most important demarcation criterion was that, to be considered scientific, theories need to be subjected to risky tests. Meehl (1978) convincingly argued typical significance tests are in no way risky tests because the null hypothesis is almost always false. *No* risk has been taken when we reject a hypothesis that we already knew was false. Lykken (1968) made a similar point: "since the null hypothesis is almost always false, one has a 50-50 chance of confirming most predictions no matter how fatuous one's theory or how illogical one's reasoning" (p.151). Lakatos (1978) also argued that NHST did not fit within a sensible model of how science progresses:

> After reading Meehl (1967) and Lykken (1968) one wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phoney corroborations and thereby a semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage (p.88-89).

Because typical NHST practice avoids subjecting a hypothesis to a risky test, the corroboration of a hypothesis it provides is very weak. It therefore fails to uphold the principles of Popperian philosophy.

Of course Oakes also raised all of these criticisms, and his use of the 'fit' between NHST and Popperian philosophy as an explanation for persistence, does not rest on that fit being real. Oakes' argument is still plausible if the fit is superficial or even completely misconceived, the misconceptions associated with *p* values being what they are. Certainly, many scientists still identify as Popperian and the (even superficial) fit of the language of NHST and Popperian philosophy is a potentially compelling appeal to authority.

Medicine, of course, has existed through the same philosophical climate as psychology (i.e., logical positivism in the 1930s, followed by Popper and falsificationism). It is not immediately obvious that this explanation for persistence

should apply only to psychology. Psychology, however, has always been, and remains, more closely affiliated with philosophy. For example, at the time of Popper's *Logic of Scientific Discovery* the disciplines of psychology and philosophy would have often still occupied a single department at many universities (Oakes, 1986). For this reason, the illusory fit Oakes discusses may indeed have had more sway in psychology than medicine, and therefore go some way to explaining the difference in reform status of the two disciplines.

7.3 Escape From Freedom

In a more recent attempt to explain the persistence of NHST in psychology, Thompson (1999) drew on existential philosophy and described "researcher resistance"[25] as an atavistic response to escape freedom and responsibility (p.135). Researchers escape responsibility, he argued, by substituting statistical significance for theoretical or practical importance and "thereby finesse the responsibility for and necessity of declaring and exposing to criticism the personal or societal values that inherently must be the basis for any decree that research results are valuable" (p.135). They simultaneously escape freedom—the freedom to analyse, judge and describe their work in a personally meaningful way. Instead they choose conformity.

Thompson cites Hagen's (1997) defence of NHST as an example of giving up on freedom: "It is unlikely that we will ever be able to divorce ourselves from that [NHST] logic even if someday we decide that we want to" (Hagen, p. 22, cited in Thompson, 1999, p.135). Further, the pressure to produce results in NHST format in order to publish has worked to enforce that conformity and further absolve researchers of both freedom and responsibility—recall, for example, Loftus' thwarted early attempts to publish without *p* values (see Chapter Four).

As for Oakes, it is easy to agree with Thompson that such factors have played a role in psychology's prolonged attachment to NHST. But again there is the question of how medical researchers overcame the problem. Perhaps in medicine, researchers were able to change practices without this impacting on their perceived sense of freedom

---

[25] I have sometimes spoken of a 'passive resistance' to statistical reform, but find 'resistance' on its own too strong a description of the state in psychology. To be sure, researchers in the main have not changed their practices in response to critics. Resistance, however, suggests something stronger than simply failing to act—it suggests a struggle. Whilst there have been some defences of NHST, reformers are for the most part simply neglected. I'm not convinced that neglect constitutes full scale resistance.

and/or responsibility? This is perhaps not implausible. Despite changes to reporting practice, it is clear from medical researchers' interpretation of their results that they have not taken responsibility for more substantive interpretations of their data. Many medical researchers have clearly not embraced their new found freedom and responsibility! Acknowledging freedom and responsibility may well be what gets medicine further down the reform road than they are now, but it does not contribute much to our understanding of why psychological researchers have not at least changed reporting practices.

7.4 Statistics as Rhetoric

In a fascinating and relatively obscure article in *Australian Psychologist*, John (1992) made perhaps one of the most sustained attempts to answer the question of why confused and/or inappropriate use of NHST persists in psychology. John cited Oakes (1986), Carver (1978), Dar (1987), Gigerenzer and Murray (1987) and several others as evidence that most explanations of the persistence had been sought in terms of individual cognitive processes (misconceptions and bias) and acknowledged that "it is not surprising that psychologists should seek explanations in terms of individual psychological process" (p. 144). Without dismissing that the misconceptions and biases identified by these authors play a role, John's own argument for persistence is qualitatively different. The 'cognitive processes explanation', he claimed, does not account for the "focal position which inferential statistics have assumed in the process of psychological knowledge production" (p.146). He proposed a more complex answer, which involves understanding the rhetorical role of NHST in psychology's appeal to epistemic authority.

John draws on the work of the French sociologist of science, Pierre Bourdieu, who viewed science as a struggle for epistemic authority. In the physical sciences such authority can be won relatively independently of a wider community or social recognition, through direct and controlled demonstration of phenomena. John quoted Bourdieu's explanation of why the same authority cannot be independent of this wider context for the social sciences:

> The power which is at stake in the internal struggle for scientific authority within the field of the social sciences, i.e. the power to produce, impose and inculcate the legitimate representation of the social world, is

> one of the things at stake in the struggle between the classes in the political field. It follows that positions in the internal struggle can never attain the degree of independence in relation to positions in the external struggle which is to be found in the natural sciences (Bourdieu, 1975, p. 36, cited in John, 1992, p. 146).

Again, the natural sciences have at their disposal something the social sciences do not—the ability to *demonstrate* knowledge:

> In the physical and natural sciences, occurrences such as space voyaging and splicing new genes into the genome of an organism are spectacular demonstrations of knowing… by comparison… the events which psychologists can bring under control, such as the serial position effect in verbal rote learning, are likely to seem very modest indeed (John, p.146).

The consequence of this, for John (and for Bourdieu) is that the production of knowledge in psychology and social science is both more open to contestation and more reliant on the discursive techniques of "negotiation, argument and persuasion" (p. 146). Such techniques become weapons in the battle for epistemic authority to be fought in the context of a wider community.

John defined argumentative persuasion or 'rhetoric' as an "informal type of reasoning for arriving at, or justifying, conclusions" (p. 147). "The appeal to scientific method", he continued, "is the major persuasive tactic used in psychological argument in the struggle for epistemic authority" (p.147). Of course, the demarcation of science and non-science—and consequently the question of what *the* scientific method is—is itself one of the most contested issues in the philosophy of science. It is relatively unsurprising, then, that NHST (with its illusory fit with Popperian falsificationism) is so easily substituted for the scientific method in this rhetorical appeal for authority.

So far John's assessment may not appear to offer more than what Oakes (1986) or Gigerenzer (1987) did before him. Yet by making explicit the ways in which NHST is used in the production of knowledge in psychology, John offers a different slant on the question of what might replace it. The revised question may ask: "what can replace the *role* of NHST in psychology's struggle for epistemic authority?" John wrote: "The illusory prospect of indubitability and conclusiveness is likely to continue to be more attractive than the cognitive discomfort occasioned by confronting the abiding uncertainty of our knowledge claims" (p. 148). As I have already said, in evaluating the benefits of CIs, they make uncertainty explicit. There is no opportunity for simply

declaring a finding 'statistically non-significant' if the interval for it is so wide as to take up the entire measurement scale. Such intervals (or at least ones wide enough to approach such a situation) may not be uncommon in psychological research. They are therefore likely to be a source of embarrassment to their authors, and not at all compelling rhetoric for their knowledge claims.

7.5 New Explanations for Disciplinary Differences

As we have seen, widespread misconceptions about NHST are not particular to psychology—they are common to medicine as well. Further, there is no reason to expect medical researchers are generally less resistant to change, or any less 'creatures of habit' than their psychology counterparts. There are other similarities between the disciplines too. First, critics in both medicine and psychology made more or less the same arguments against NHST. Second, in both disciplines, they (mostly) made them clearly and eloquently. Third, critics were published regularly in respected journals, that is, they were not marginalised. Fourth, both disciplines had some editors committed to improving statistical reporting practice in their journals.

To some extent, the lag is psychology can be accounted for by historical accident. Arguably the most influential reform advocate in psychology, Jacob Cohen, died during the tenure of the TFSI. In fact, he died some time before the subcommittee for revising the APA *Publication Manual* was even established. Furthermore, Robert Abelson, a co-chair of the TFSI, became very ill with Parkinsons disease during the TFSI tenure. This left Robert Rosenthal as the only chair. The TFSI was in a sense not directed by (the majority) of those intended to lead it. Such contingencies need to be acknowledged, even though their influence is difficult to measure. Oakes, Thompson and John all provide accounts of the persistence of NHST in psychology goes some way to providing a sociological explanation of the persistence of flawed practices in this discipline; the factors listed in the next section offer perhaps more specific reasons for the differences in reform progress in the two disciplines.

The Nature of Editorial Policy: Requirements Vs Encouragements

Janet Lang (a former co-editor of *Epidemiology*) proposed that the difference between successfully reformed and less successfully reformed medical journals could be explained by how well the policy relating to CIs was enforced (personal

correspondence, May 2003). For example, at *Epidemiology* there was greater success in instituting CIs (and to some extent removing binary decisions based on *p* values), because of the strict enforcing by Rothman, Lang and their co-editors. On the other hand, CI reporting was not as common in the *American Journal of Epidemiology*, because the policy had not been enforced (Savitz, Tolo & Poole, 1993).

Lang extended this argument to account for why reform had been unsuccessful in psychology journals such *Memory and Cognition* and the *Journal of Consulting and Clinical Psychology (JCCP)*. She wrote:

> It seems to me that this issue of policy enforcement is key. It is not the author's fault if the journal does not enforce its own policy. Having a policy that is declared but obviously not enforced might even weaken the odds of changing the behavior in question (personal correspondence, May, 2003).

Her argument certainly seems plausible. Compare, for example, the editorial position of Ken Rothman (*Epidemiology*) with that of Philip Kendall (*JCCP*). Rothman's editorial stated: "In *Epidemiology* we do not publish them [*p* values] at all." (1998, p.9). The proof of its impact: In 2000, *Epidemiology* did not publish a single *p* value, and 94% of empirical articles reported CIs (Fidler, Thomason, Cumming et al., 2004). Kendall, on the other hand, wrote: "Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance" (1997, p.3). Just 40% of authors followed this *encouragement* to report clinical significance (Fidler, Thomason, Cumming et al., 2004). Unlike Rothman, Kendall 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" (Philip Kendall, personal communication, April 9, 2001).

Rothman's policy was extreme even for medicine. Shrout (1997) called it a "virtual ban" (p. 1). However, it was not only Rothman's policy that could be considered more strict than most policies in psychology. For example, Langman (1986) at the *British Medical Journal (BMJ)* was also direct: "…from 1 July authors of papers submitted to the *BMJ* will be expected to calculate confidence intervals whenever the data warrant this approach" (p. 716). As I pointed out in Chapter Six, this language of 'expectation' was reinforced by those that provided practical instructions and guidance: "The *British Medical Journal* now expects scientific papers submitted to it to contain confidence intervals…" (Gardner and Altman, 1989, 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* is the only exception to this I know of. Suggestions of requirements, bans or mandates related to statistical reporting have largely been met negatively. The original proposal to ban *p* values was quickly dismissed by the TFSI; the APA *Publication Manual* committee did not even broach the question. Even some advocates of reform in psychologist I interviewed described the idea of bans or requirements as impinging on researchers' intellectual freedom. The other side of the argument—that is, the argument for requirements—is perhaps presented best by Bruce Thompson. To requote:

> 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 editorial reform processes. 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 (Fleiss, 1986; Shrout, 1997). But as Rothman explained, for him reform was not an issue of intellectual freedom, just correcting mistakes as one would correct grammatical errors.

Geoff Loftus has arguably been one of the strictest editors psychology has seen on this matter. It is worth considering, then, why his policy was not enforced. Recall that it clearly it wasn't: Over half the papers published during his term failed to follow his recommendations (Finch, Cumming, Williams et al., 2004). Why would someone so committed to reform not enforce their policy? The answer is that Loftus encountered considerable resistance to his policy—well beyond the sort of complaints Rothman received from Fleiss (1986). Authors simply failed to provide the intervals Loftus requested. As I explained in Chapter Four, Loftus worked hard to counter this resistance, personally calculating around 100 standard errors and CIs for authors who did not provide them. Rothman, on the other hand, recalls encountering little resistance in enforcing his policy, and certainly did not have to do these sorts of calculations himself (Ken 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? The remainder of the sociological factors listed in this chapter address this question.

The Importance of Re-writing Textbooks

In the 1980s Gardner and Altman identified an obstacle to the success of editorial reforms in medicine: "…the methods needed to calculate confidence intervals are not readily available in most statistical textbooks" (1989, p.4). They addressed this problem directly by writing such a textbook, *Statistics with Confidence*, which included dedicated software, *Confidence Interval Analysis*. Rothman (1988) offered *Modern Epidemiology*, an advance statistical text that supported his CI approach to analysis. Both texts are now in second editions (Altman et al., 2000; Rothman & Greenland, 1998). In psychology the equivalent texts have only been published in the past few years (e.g., Kline, 2004; Lockhart, 1998; Smithson, 2002; Thompson, in press; Zechmeister & Posavac, 2003). However, such texts are not necessarily widely adopted: "Unfortunately it's not selling well at all… Moreover, as far as I know no review has been published… Even instructors who have a basic sympathy for the approach seem reluctant to undertake what they see as a radical change to the way they teach." (Robert Lockhart, personal correspondence with an anonymous examiner of this thesis, 2000). Surveys of teaching curricula and textbooks in psychology have also shown few signs of change (Aiken, West, Sechrest & Reno, 1990; Azar, 2001).

The Need for Editorial Collaborations

Members of the International Committee of Medical Journal Editors (ICMJE, introduced in Chapter Six) met annually during the mid 1980s, and a focus of these meetings was statistical reporting in their journals. It was through these meetings that the editorial reforms initiated in journals like *American Journal of Public Health (AJPH)* and *BMJ* spread to other journals. For example, as a result of ICMJE consultations the then editor of the *Medicine Journal of Australia (MJA)* wrote an extended editorial on the benefits of CIs over statistical significance (Berry, 1986; Geoff Berry, personal correspondence, April 2003). By the late 1980s, virtually all the major journals—*NEJM, BMJ*, *AJPH, Lancet*, *British Heart Journal*, *Circulation Research*, *JAMA* and others—had policies recommending CIs, backed up by the ICMJE guidelines.

In psychology, on the other hand, Geoff Loftus was attempting to enforce his policy on error bars at *Memory and Cognition* in 1993—three years before the TFSI was even conceived, six years before their guidelines were published and eight years before the APA *Publication Manual* included a CI recommendation. As a single editor Loftus had only a limited and short-lived impact.

We might expect things to change for psychology when APA came on board. But even then their collaboration with editors was minimal. Some TFSI members were themselves editors—for example, Mark Appelbaum was editor of *Psychological Methods* and Bruce Thompson was editor of *Educational and Psychological Measurement.* As I have explained, however, there was no systematic attempt to consult with other editors of APA, APS or independent psychology journals throughout the process. The TFSI eventually addressed a meeting of APA editors, shortly after their 1999 guidelines were published. By this time, the TFSI tenure had virtually ended and the group was on the verge of disbanding. In my interview with Robert Rosenthal, he recalled the meeting of APA editors at which he, as co-chair of the TFSI, addressed the editors. The meeting was largely unsuccessful from Rosenthal's point of view. There was little reaction of any kind to the newly published TFSI guidelines from editors and no follow-up meetings. In fact, Rosenthal described it as "a depressing experience" (Robert Rosenthal, personal communication, May 2003).

The Role of Statistical Editors and Reviewers

Marks, Dawson-Saunders, Bailar, Dan and Verran (1988) wrote: "the increased involvement of statisticians in the publication of biomedical research has resulted in increased communication between statisticians and biomedical journal editors" (p. 1003). Many medical journals have statistical editors as well as substantive editors, and the scope of statistical reviewing is wide (George, 1985; Altman, 1991). This is not the case in psychology. Editors of substantive journals in psychology are generally not statisticians. They are leading researchers in development, social, clinical, industrial or what ever other are of substantive psychology. They are consequently, and somewhat justifiably, not necessarily concerned with what they consider to be 'not their game.'

How it came to be that medicine adopted statistical editors and reviewers and psychology did not is one of the complex motivational questions that I make no claim to provide a comprehensive answer for. My hypothesis, for what it is worth, is again an economic one. First, the far greater cost of medical research motivated higher levels in

caution than in psychology and, second, that medical journals simply have larger budgets for organising associate editors. Psychology internalised the responsibility for statistical analysis, whereas medicine appropriated outside expertise.

The Development of Methods Journals

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 USA Psychology departments, only content-based articles on a substantive (e.g., clinical, development, social) research topic were counted as professional publications in tenure applications (Roger Kirk, personal communication, September 2001). Articles published in statistics journals were not rated, even when they were primarily about methodology and statistics in psychological studies. *Psychological Methods* provided—for the first time for staff in departments affected by such regulations—an opportunity to publish on statistics and methodology, without being penalised by tenure and promotion processes.

How widespread were the effects of this journal being launched? It is difficult to say what impact this would have had on the output and productivity of quantitative psychologists. Other statistical journals ostensibly within the field of psychology (such as *British Journal of Mathematical and Statistical Psychology*, *Psychonomic Bulletin and Review* and *Behaviour, Research Methods Instruments and Computers*) are listed under Mathematical Psychology in Journal Citation Reports (2004) whereas *Psychological Methods* is in the general listing. Not surprisingly its impact factor is much higher than any of the 10 journals in the Mathematical Psychology listing: *Psychological Methods* has an impact factor of 5.525 whereas the highest of the Mathematical Psychology listings, *Psychonomic Bulletin and Review*, is 1.931. This in itself would have implications for not only the readership articles would get, but also, under some systems, the benefits (e.g., increasing tenure prospects) publishing researchers receive for their work.

The Benefits of In-house Statisticians

Virtually all medical schools employ at least one statistician in a consultative role—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, statisticians aligned themselves with advocates of clinical trials. Marks (1997) claimed that since the 1960s, medical researchers have more often consulted with statisticians about the design of studies (although many would argue that the rate of pre-data collection consultation should be higher). In psychology, such consultation is rare. Psychology departments virtually never employ in-house statisticians for consultation. It is again hard to provide evidence of changes in consultation rates, or their impacts. If Marks is right, however, the implications of this difference between the disciplines are likely to be widespread and important.

7.6 More Conceptual Matters

Medicine's shift was from statistical significance to effect estimation. The same shift has been proposed for psychology—but is estimation as the basis of psychology conceivable? One reason it is difficult to imagine an estimation-based psychology is, in many areas, 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 often universal. For example, everyone measures blood pressure in millimetres of mercury[26].

In psychology, on the other hand, one study might measure anxiety using a particular anxiety inventory or test; another might measure it with a different inventory or by increases in heart rate or skin conductance. How can these studies ever be compared? Such judgements are essential to the growth of cumulative knowledge and the progress of science, and are not necessarily straightforward.

Increased reporting of standardised effect sizes and meta-analysis have been strongly advocated as a solution to the problem of comparison (Hunter & Schmidt, 2004; Thompson, 2002). Jacob Cohen explained that standardised effect sizes were necessary in psychology because "the units in which we measure our dependent variables are not only arbitrary but also without any absolute meaning, not at all like inches or pounds or degrees Fahrenheit, which we have all experienced in many contexts" (1965, p. 102). However, Oakes (1986) disputed Cohen's claim that this problem was unique to psychology.

---

[26] It would be dangerous to generalise too far here; blood pressure is a very convenient example. Sometimes the measurement scales in psychology and medicine would overlap, with psychology using medical outcome measures and medicine using psychological outcome measures.

> But all constructs, whether they be distance, weight or temperature on the one hand, or anxiety, alienation, or reaction time, on the other, are man-made. The natural sciences cannot be differentiated from the social sciences on the basis of the extent to which 'we have experienced' their constructs. Indeed, a great many psychological constructs have derived from common experience, the same cannot be said, surely, of osmotic pressure, electrogmagnetism, super- conductivity, or electron orbits. (p.62).

Oakes misses something in his analysis of this problem I think. It is not only naturalness or meaningfulness that distinguishes medicine from psychology, but often simply the level of consensus over the scales on which concepts will be measured—in other words, the universality of measurement scales. Further, in medicine it is more common to employ scales which have a rational zero point (ratio scales); in psychology, we more commonly see scales which lack a zero point (interval scales). It is not simply that psychology does not have ratio scales; in many cases it is difficult to see how they could even be applied to psychological constructs. What would zero intelligence mean? Or zero extraversion? Any ratio scale for a measure can be easily converted into any other ratio scale for the same measure, but this is not the case for interval scales. Medicine had no reason to employ standardised effect sizes for the purposes of comparison or meta-analysis, because of the existing universality of scales and/or rational zero points. In medicine, CIs have therefore overwhelmingly been applied either directly to the raw measures (means, percentages) where consensus is high, or to other common units-free (but not standardised in the sense of being divided by the standard deviation) measures such as odds ratios (OR) and relative risks (RR) where these measures are more appropriate. Even when medicine relies on psychological outcome measures, there may be greater consistency in use of the measures because clinical trials often focus on predefined primary outcomes.

If standardised effect sizes are the way forward for psychology, the application of CIs will not be as straightforward as it was for medicine's chosen effect sizes. CI calculations for standardised effect sizes typically require non-central distributions. Few psychologists have any experience in working with standardised effect sizes; they are certainly not part of any mainstream psychology training. The calculations themselves involve iterative algorithms and scripts for running such processes that have

only been published relatively recently and are still relatively obscure (Cumming & Finch, 2001; Fidler & Thompson, 2001; Smithson, 2001; Steiger & Fouladi, 1997).

How important is it to use a non-central distribution for these effect sizes? When sigma is *known*, for example, in the case of IQ, *d* will be normally distributed and a non-central distribution irrelevant. When sigma is *unknown* and when the null is true, central *t* also applies. However, when sigma is unknown and the null is false, the non-central distribution becomes important. As Kelley (2005) explained:

> …when the null hypothesis is true, the difference between two group means is normally distributed about zero. When this difference is divided by its standard error it follows a central *t* distribution… However, when the null hypothesis is false, the difference between the two means divided by its standard error does not follow a central *t* distirubiton rather it follows a nonsymmetric distribution that is known as noncentral *t*… (p. 53).

The non-centrality parameter is a function of the population standardised effect size and the sample size. The degree to which using a non-central distribution makes a difference to the confidence limits depends on the size of the non-centrality parameter. In other words, the answer to "does it make a difference to use a non-central distribution?" is that it depends of the size of the effect and the sample size. For very large effects or very small samples, the difference may be a substantial overestimate or underestimate of the imprecision. The rest of the time, using a non-central distribution may not produce results drastically different to bootstrap methods or even regular parametric methods. Kelley (2005) compared bootstrapped CIs and non-central CIs for *d,* under various conditions. In all cases "the non-central method outperformed the bootstrap" (p.55) methods but in practical terms, the bootstrap procedure[27] often works just as well.

The statistical reform hurdle 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. Second, once we move to the realm of standardised effect sizes, calculating CIs becomes a more

---

[27] Kelley tested two bootstrap procedures: a bootstrap percentile method and a bootstrap bias-correct (to correct the bias of the *d* statistic) and accelerated method. The later performed much better than the former, and produced results closest to the non-central method. Kelley recommends against the percentile bootstrap method.

complex enterprise (at least in those cases where non-central distributions will make a substantial difference) and the intervals themselves become harder to interpret.

Neither of these problems was encountered in medicine to anywhere near the same degree as in psychology. Medicine is often a practical, utility-based discipline. The conceptual shift to estimation would have been relatively straightforward. Where effect measures were artificially constructed and not inherently meaningful, the universality of measurement relieved serious difficulty. Although they are units-free, OR and RR values still rely on the original scale of the study for full interpretation. Their calculations do not generally require the added complication of non-central distributions, as standardised effect sizes do. Some reform advocates in medicine have gone further, arguing that standardised effect sizes are not only unnecessary in medicine, but in fact are invalid (see following section).

Along with many expository journal articles, texts such as Gardner and Altman (1989) and Rothman (1988) provided guidelines for CI calculations for medicine appropriate effect sizes. Comparatively, psychology's attempts to provide comprehensive CI methods for the effect sizes advocated were very late, and as I have said, still comparatively obscure.

Of course, *all* the difficulties for psychology raised here by the proposed shift to estimation still exist for this discipline within a hypothesis testing paradigm. The point is that a shift to estimation exposes them, forcing psychologists to confront some serious, fundamental and potentially uncomfortable philosophical questions about the enterprise they are engaged in.

The Standardised Effect Size Debate

Standardised effect sizes have been heavily criticised by some medical researchers and statisticians, on the grounds that they compound uncertainty. Cohen's *d*, for example, divides an estimate of the mean, which has error, by an estimate of the standard deviation, which also has error. Sander Greenland has been the most outspoken critic of these effect measures. It is difficult to gauge how widespread Greenland's view is; there is little reference to the problem in other medical literature. What is clear is that standardised effect sizes are not *used* in this discipline. In our analysis of *AJPH* and *Epidemiology* we did not find a single instance of such an effect size in any of the 700 articles surveyed. Whether this is because researchers simply find them unnecessary or whether they believe them to be invalid is not a question I can

answer based on our study. Greenland, Schlesselman and Criqui (1986) explained why the standard unit is deceptive:

> …the 'standard unit' at issue here is merely the sample standard deviation of the study, and this quantity in no way conforms to ordinary English or scientific concepts of 'standard unit': after all, the standard deviation can vary dramatically upon changing the study design, the variable under discussion, or the target population. (Contrast this statistical use to use of "standard unit" such as an international unit of Vitamin A.) We think it is simply misleading to term something a 'standard unit' when it is in fact a high variable quantity" (p. 207-208).

Greenland further argued that "standardized regression coefficients, correlations, and 'variance explained' are also improper summaries of effect." (1987, p.767) Again, his criticism is that these measures "confound the effect of a risk factor with the standard deviations of the factor and the disease" (1986, p.203). Greenland holds partially responsible a general confusion of *variation* and *variance* for perpetuating what he believes to be misguided advocacy of standardised measures.

If one avoids standardised effect sizes and coefficients, as Greenland suggests, how are results across studies to be compared (in the absence of meaningful or universally agreed upon raw units)? To his credit, Greenland does not dismiss this question. He offered the following scenario: Imagine we wanted to compare the effects of low density lipoprotein (LDL) cholesterol (measured in mg/dl) versus blood pressure (measured in mmHg) on coronary heart disease. His advice on how to proceed follows:

> …there is no definitive method for such comparison… there may be some validity in comparing estimates of the effect based on reference and index values that are 'natural' or biologically meaningful… For example, one could estimate the increase in LDL cholesterol necessary to produce a 50 per cent increase in coronary heart disease risk and compare this with the estimated increase in blood pressure necessary to produce the same increase in risk (1986, p.207).

For psychology, the equivalent to Greenland's suggestion might be a protocol for reporting measures of clinical significance. Certainly there has been some recent progress in developing such measures, such as the Reliable Change Index and Normative Comparisons. As I mentioned in Chapter Four, *JCCP* ran a special section on clinical significance, including how to calculate a variety of these measures, in 1999.

What remains absent from psychology, as far as I am aware, is discussion about how these clinical significance measures might be used to compare studies and to conduct meta-analysis.

It is important to point out that Hunter and Schmidt outline techniques for overcoming the difficulty of standard deviation not being a 'standard unit' (see Hunter and Schmidt, 2004[28], chapters 3 to 8). *Methods of Meta-Analysis* devotes many pages to this issue, explaining techniques for correcting not only systematic problems (measurement error, range variation), but also sampling error, for correlational effect sizes and for *d*. With such corrections, standardised effect sizes can play a crucial role in the cumulation of scientific knowledge. In many cases however, standardised effect measures are often recommended by statistical reformers and advocates of meta-analysis, without any hint at the criticisms that have been made or the corrections that may be relevant or needed.

7.7 A Broader Problem

The failure to provide adequate CI instruction and guidance for recommended effect sizes can be seen as part of a broader problem with statistical reform in psychology. Reform advocates in psychology have from time-to-time been admonished for relying on experimental scenarios that are over-simplified. To requote Grayson, Pattison and Robins (1997): "some recent attacks on significance testing in the psychological literature… have largely taken place in the context of simple models with few parameters." (p. 69). In Faulkner, Fidler and Cumming (2005) we demonstrated that in clinical psychology at least, designs are indeed complicated—with an average of 13 dependent variables in articles surveyed. If Grayson, Pattison and Robins are right, then reformers themselves must also be held responsible for the lag in psychology. The good news for ecology is that this criticism seems not to apply: Information theoretic and Bayesian approaches have largely been advocated in the context of complex, real world problems (see Chapter Eight for statistical reform in ecology).

[28] These techniques are also outlined, in the same chapters, in the 1990 edition of *Methods of Meta-analysis*.

7.8 Summary

Improvements in statistical reporting in medicine, specifically the move to CIs, can be partially attributed to strict editorial policy: In journals where policies were strictly enforced, the changes were most dramatic. That editorial reforms in leading journals were virtually simultaneous seems important, and is certainly a distinguishing factor of medicine. The timely re-writing of statistical textbooks (to fit with policy recommendations) was also particular to medicine and institutional support and guidelines from the ICMJE coincided with both editorial reform and the rewriting of textbooks. Conversely, psychology's lack of reform might be explained by editors working in isolation, a lag in the re-writing of textbooks and inconsistent and delayed advice from the APA *Publication Manual*. In addition, the transformation to a science of estimation is perhaps itself a more difficult task for psychology, conceptually and computationally.

One reason editorial policy changes have been, and will continue to be, insufficient for full reform, even in medicine, 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. How are CIs 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. What has so far been conspicuously absent from reform debates, in any of these disciplines, is an evidence-based approach—a topic I address again in Chapters Nine and Ten. One can only hope that when reform does occur in psychology, 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.

8
STATISTICAL REFORM IN ECOLOGY

In my view, NHST has done more damage to the environment than a veracious, unprincipled mining or logging company (Mark Burgman, personal communication, November 2005).

Many wildlife biologists and ecologists have changed their perspectives regarding data analysis as a result of the limitations of null hypothesis testing. Some investigators have merely refocused attention on estimating effect size and measuring their precision, without undue emphasis on a null hypothesis, test statistics, P-values, and arbitrary notions of significance… This simple approach is effective for many biological questions. We support the estimation of effect size and use it in our own research work. Some investigators have begun to explore an assortment of Bayesian methods… Others have begun to use the relatively new information-theoretic methods…(Anderson & Burnham, 2002, p. 912).

Statistical reform efforts have a much shorter history in ecology than in other disciplines—and have come from more directions. Early warnings about the consequences of ignoring statistical power began in the 1980s, but few authors provided any broader criticisms of NHST or advocated alternatives. More recently, advocates of Bayesian methods have been amongst the most outspoken critics (Ellison, 1996; Harwood, 2000; Wade, 2000; Clark & Lavine, 2001). As I have mentioned, so too have proponents of likelihood and information theoretic methods (particularly, Anderson et al, 2000; Burnham & Anderson, 2001), with Akaike's Information Criteria (AIC) receiving particular attention.

Unlike medicine and psychology, there has been relatively little attention paid to CIs in ecology (with the exceptions of Cherry, 1996, 1998; DiStefano, 2003). Whilst the criticisms of NHST are now virtually identical to those made in medicine and psychology, in other respects the approach taken to statistical reform has been quite different.

In ecology, Bayesian and information theoretic approaches will no doubt grow in popularity as their computational difficulties become less daunting, with faster computers and better-developed software. However, they are yet not incorporated in most undergraduate curricula and mainstream training of ecologists. As a journal survey presented in this chapter demonstrates, the research literature remains dominated by NHST but nascent signs of change can be detected.

8.1 Reform from the 1980s: The Statistical Power Debate

Statistical reform literature in ecology during the 1980s and early 1990s focused almost exclusively on the issue of statistical power. Calls for increased attention to statistical power began in the early 1980s (e.g., Berstein & Zalinski, 1983; Mapstone, 1985; Toft & Shea, 1983; Vaughan & van Winkle, 1982). As in other disciplines, they were largely unsuccessful. For example, Peterman (1990) surveyed 1987-1989 issues of the *Canadian Journal of Fisheries and Aquatic Sciences* and the *North American Journal of Fisheries Management* and found little response to be found to earlier critics. In fact, out of 408 papers, he found only 3 instances of statistical power being reported! Yet, of the 160 papers that failed to reject the null hypothesis, over half (52%) made an assertion of 'no effect' (despite the mentioned virtual absence of any power information). Later surveys found still no response. For example, Thomas and Juanes (1996) reported than in 359 research articles in 1994 issues of *Animal Behaviour* "279 included at least one statistically non-significant result but only one… reported the power of the tests used" (p. 859).

Lack of consensus in the literature about how to properly calculate and interpret statistical power, may be partially responsible for early failure of power advocates to reform practices. In ecology power advocates often failed to distinguish between: a) statistical power calculated *a priori*, b) statistical power calculated retrospectively using the expected effect size, and c) statistical power calculated retrospectively using the obtained effect size. The first is of course preferable, and the third is almost entirely useless. Yet, as I explained in Chapter Two some ecology articles in the mid 1990s explicitly recommended this third practice (Reed & Blaunstein, 1995; Thomas & Juanes, 1996). In fact, between 1983 and 1997, Hoenig and Heisey (2001) identified 19 independent articles[29] advocating "post-experiment power analysis" (p.20). A hot debate over the utility of retrospective power analysis based on the observed effect size followed. The practice has now been severely and justifiably criticized although it remains difficult to estimate how widespread the remaining confusion over this issue might be.

---

[29] Two articles were from psychology journals and a third from an education journal. The others were ecology or biology related.

Unlike psychology and medicine, reform in ecology seemingly moved straight from this debate over statistical power to discussion of information theoretic and Bayesian methods—with little time or attention given to CIs along the way. Perhaps this will serve the discipline well. For the moment, it is too early to say.

8.2 Editorial Policy

In ecology, there have to date been fewer attempts to improve practice through editorial policy than in medicine or psychology. Below are the only examples I could find.

The Wildlife Society Journals

Two of the Wildlife Society's journals, *The Wildlife Society Bulletin* and the *Journal of Wildlife Management*, have gone some way to addressing statistical reporting problems. The *Journal of Wildlife Management (JWM)* published a 1995 editorial encouraging statistical power and, unusually for ecology, the use of CIs. This journal has been particularly active in statistical reform, regularly publishing related articles, including some of the most well-known articles on this topic in the ecological literature (e.g. Johnson, 1999; Anderson et al., 2000). Unfortunately, the 1995 editorial bordered on misinterpretation—both in its ambiguous description of statistical power and in its attempt to provide guidelines to interpreting CIs. For example, on interpreting CIs, the editorial stated: "The smaller the confidence interval (or detectable difference), the more biologically meaningful the reported result" (*JWM*, 1995, p. 197). At a stretch, this can be read as a narrower interval means the study is more precise and the results therefore more focused. However, it is a shame that the one of the only journals to make such an editorial effort is riddled with such ambiguities. Some problems in the editorial were corrected in a letter to the editor in a later issue of the journal (Otis, 1995). For example, this correction was made to the advice about statistical power:

> [the editorial] implies that there are 3 factors that determine test power: Type I error, sample size and effect size, which is described as 'the difference between 2 samples that is of biological importance'... This description gives the reader the impression that population variance and experimental error is not a factor in power analysis (p. 630).

A second journal of this society, *The Wildlife Society Bulletin*, published an article by Steve Cherry in 1998, calling for The Wildlife Society to "explore ways to help authors and reviewers conduct statistical analysis more effectively…Editors and associate editors of the *Journal* and the *Bulletin* could help by making it clear that *P*-values are not a prequisite for publication." (p.852).

Despite Cherry's challenge to the editors, the *Bulletin*'s current guidelines to authors contain little in the way of statistics recommendations. In fact, one of the recommendations is almost a reconstruction of a typical NHST misconception: "Avoid redundant use of the word 'significantly' (e.g., 'the means differed [p=.016]')" (Andrews & Leopold, 1999[30], p. 10)—in other words, it allows (more-over recommends) 'statistically significant difference' to be substituted for 'difference', and presumably 'statistically non-significant difference' to be substituted for 'no difference'. Unfortunately, the *Bulletin's* guidelines also indirectly endorse the use of 'asterisks significance' and the interpretation of statistical non-significance as 'no effect', as shown in the following example. When explaining the correct use of superscript in footnotes, the example reads: "Use Roman uppercase letters instead of rules (e.g., 12.3Aa, 16.2A, 19.5B) where the superscript 'a' references a footnote such as 'Means with the same letters are not different ($P < 0.10$)'" (p. 17).

The Ecological Society of America Journals

The Ecological Society of America (ESA) publishes 6 journals including *Ecology*, *Ecological Applications* and *Ecological Monographs*. Their 'Guidelines for Statistical Analysis and Data Presentation' (http://www.esapubs.org/esapubs/Statistics.htm, last cited 07-11-05) go some way towards encouraging reform. For example, they: point out that effect size and biological importance should not be confused with statistical significance; recommend including a measure of precision (such as a standard error or CIs); and encourage graphical presentation of results. Unfortunately, CIs are misleadingly referred to as "descriptive procedures."

Not unlike the fifth edition of the APA *Publication Manual*, the main problem with the ESA guidelines is not in the recommendations themselves, but rather in the lack of 'follow through'. There are no examples of how to institute what might be

[30] Andrews & Leopold (1999) are still the guidelines currently offered to authors in this journal.

unfamiliar practices, or to indicate best practice reporting. This is seemingly tied up, as it was for the *Manual*, in the reluctance to make *requirements* for statistical reporting. For example, the ESA guidelines describe their basic philosophy as quoted below. The first point is most important to the current argument, and parallels sentiments expressed by the APA *Manual* committee to not impinge on authors' intellectual freedom:

> Authors are free to perform and interpret statistical analyses as they see fit.
>
> The reader needs to be provided information sufficient for an independent assessment of the appropriateness of the method. (http://www.esapubs.org/esapubs/Statistics.htm)

The problem with this is, as Bruce Thompson noted, to present an *encouragement* in the face of other *requirements* sends a self-cancelling message. In psychology, this is particularly problematic because of the myriad requirements, in the case of the *Manual* about margins, formatting, pagination etc. In ecology, it is not clear that as damaging a message is sent by philosophy outlined by the ESA. This is because the *encouragements* they provide are at least not made in the face of the same myriad *requirements*. Within reason, matters of style are flexible in this discipline—at least more flexible than they are in psychology. In other words, although the ESA guidelines may to some extent mirror the APA *Manual's* sentiments, it is not necessarily the case that they will have the same negative, self-cancelling impact on researchers' statistical practices.

8.3 Have Criticisms of Null Hypothesis Significance Testing Had an Impact on Statistical Reporting Practices in Conservation Biology?

I narrow the scope of this chapter somewhat now to focus specifically on conservation biology, because conservation biology has a particular interest in getting things right: "The consequences of accepting a false null hypothesis can be acute in conservation biology because endangered populations leave very little margin for recovery from incorrect management decisions" (Taylor & Gerrodite 1993, p.489). Where populations are small "waiting for a statistically significant decline before instituting stronger protection measures" (p.493) is often tantamount to a guarantee of extinction. Incomplete statistical reporting, particularly low and unknown statistical

power, can result in direct, unanticipated and unacceptable environmental damage. As we know, in psychology half a century of criticisms, and a decade of editorial and institutional intervention, has had little impact on statistical reporting practices in journals. The primary question of this survey is has conservation biology, as a discipline, been equally resistant to change? To answer this question I surveyed articles in the two leading conservation biology journals: *Conservation Biology* and *Biological Conservation.*

8.3.1 Method

I coded statistics in 50 articles published in 2001 and 2002 in each of *Conservation Biology* and *Biological Conservation* and 50 articles published in each journal in 2005. I coded articles with empirical data only and did not code meta-analyses, or methodological or theoretical articles. In any article I coded the first occurrence of each item listed in the first column of Table 8.1. I calculated the proportion of articles reporting each item, and 95% CIs for those proportions using the method recommended by Newcombe and Altman (2000). Approximately 10% of 2001 and 2002 articles, selected to represent the full range of article types, were independently cross-coded by an ecologist. The accuracy of my coding was 92%; discrepancies were oversights, often due to unfamiliar formatting of results[31], rather than disagreements over definitions.

8.3.2 Results

Results from the survey suggest that conservation biology, unlike psychology, has demonstrated a convincing, albeit small, response to criticisms. Table 8.1 gives rates of the statistical reporting practices coded. In 2001 and 2002, 92% of sampled articles in *Conservation Biology* and *Biological Conservation* reported at least one *p* value. In 2005, this figure had dropped to 78%. In 2001 and 2002, 7 of the 8 articles without NHST reported descriptive statistics only. In 2005, of the 22 articles that did not use NHST: 7 were descriptive, 9 built mathematical models, 4 used AIC model selection techniques, 1 was Bayesian and 1 reported CIs only. In addition, of those articles that

[31] I am not an ecologist.

did report null hypothesis test results, 4 supplemented this analysis with AIC, 2 with maximum likelihood estimates and 2 with Bayesian analysis.

In 2001 and 2002, statistically non-significant results were often reported without statistical power: 80% of articles that used NHST reported a non-significant result, yet only 3% of these reported power. In 2005, there was a slight increase in the reporting of statistical power: 86% of articles that used NHST reported a statistically non-significant result, and 8% of those reported statistical power. The high reporting rate of statistically non-significant results may be good news in itself. It provides some evidence that the file draw problem may not be a serious concern for this discipline. There was also a modest increase in the reporting rate of CIs: from 19% in 2001-2002 to 26% in 2005—and in the percentage of figures using error bars (standard error or CI), from 40% to 51%. Figure 8.1 summarises the percentages of articles that included NHST, power, and CIs, in 2001-2002 and 2005.

In 2001 and 2002 statistically non-significant results were interpreted as evidence for 'no effect' or 'no relationship' in 47% of articles that included a non-significant result—despite the rarity of statistical power information. In 2005, this misconception was present in 63% of such articles. This increase can not be accounted for by the small increase in statistical power reporting. In addition, there remain many cases of *p* values being reported without effect size estimates, measures of variance or sample size information. Table 8.2 gives these reporting rates for 2000-2001 and 2005.

8.3.3 Survey Conclusions

The statistical reporting practices in two leading conservation biology journals indicate promising change. The decline in NHST reporting (from 92% in 2001 and 2002 to 78% in 2005) sets conservation biology apart from psychology in terms of reform. However Figure 8.1 illustrates the still limited extent of changes, and the continuing dominance of statistical significance testing.

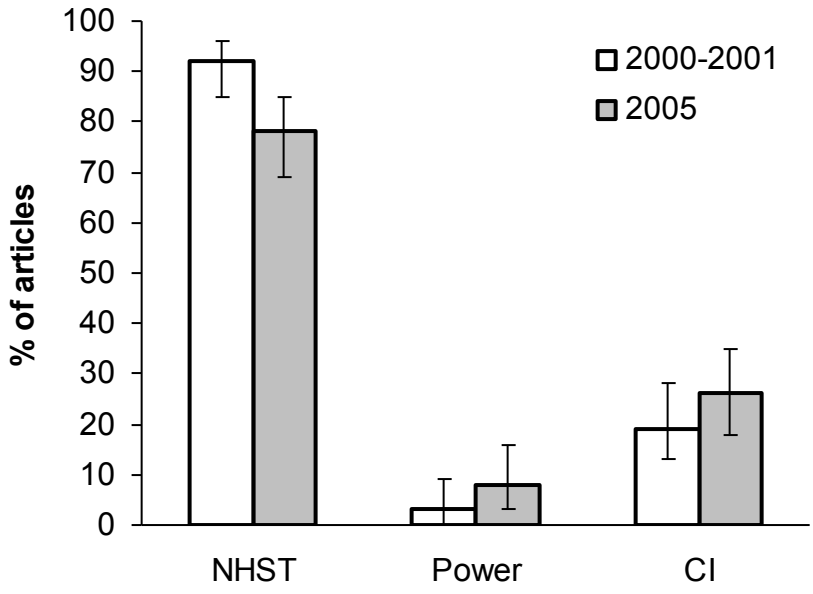


Figure 8.1. Percentages of conservation biology articles reporting null hypothesis testing, statistical power and CIs. Error bars are 95% CIs (Newcombe & Altman, 2000).

Table 8.2.
Percentage of articles with statistical significance tests that also reported, or omitted, an effect size measure, variance measure (SD or SE) or sample size (n or df). 95%CIs calculated according to method recommended by Newcombe and Altman (2000).

Item:	2000 and 2001		2005	
	% (of 92)	95% CI	% (of 78)	95% CI
With at least one effect size	87%	79 to 92%	89%	80 to 94%
Missing at least one effect size	43%	34 to 54%	58%	47 to 68%
With at least one SD or SE	48%	38 to 58%	47%	37 to 58%
Missing at least one SD or SE	67%	57 to 76%	85%	75 to 91%
With at least one n or df	76%	66 to 84%	77%	66 to 85%
Missing at least one n or df	36%	27 to 46%	51%	40 to 62%

Note: We classified the following measures as an effect size: mean (or percentage or proportion) difference; any relevant standardized measure, such as Cohen's *d*; *b* or *Beta*; Variance-accounted-for measures such as R^2 (for regression) or η^2 (for ANOVA); correlation coefficients and other unit-free measures, such as odds ratios.

Table 8.1.
Percentage of articles reporting statistical significance tests, confidence intervals and figures in conservation biology journals. 95%CIs calculated according to method recommended by Newcombe and Altman (2000).

Item:	2001 and 2002		2005	
	% articles	95% CI	% articles	95% CI
Any NHST	92% (92 of 100)	85 to 96%	78% (78 of 100)	69 to 85%
-nil null hypothesis[1]	79% (73 of 92)	70 to 86%	97% (76 of 78)	91 to 99%
-ambiguous use of 'significant'[2]	68% (63 of 92)	58 to 77%	63% (49 of 78)	52 to 73%
-exact p value[3]	62% (56 of 92)	51 to 70%	69% (54 of 78)	58 to 78%
-p value asterisks (i.e. *, **) [4]	25% (23 of 92)	17 to 35%	22% (17 of 78)	14 to 32%
-non-significant result	80% (74 of 92)	71 to 87%	86% (67 of 78)	77 to 92%
-statistical power	3% (2 of 74)	0 to 9%	8% (5 of 67)	3 to 16%
-indirect reference to power[5]	30% (22 of 74)	21 to 41%	30% (20 of 67)	20 to 42%
-interpret as 'no effect'	47% (35 of 74)	35 to 57%	63% (42 of 67)	51 to 73%
Any CI	19% (19 of 100)	13 to 28%	26% (26 of 100)	18 to 35%
-interpret CI[6]	26% (5 of 19)	12 to 49%	31% (8 of 26)	17 to 50%
Any figure with data	77% (77 of 100)	68 to 84%	69% (69 of 100)	59 to 77%
-error bars on figure[7]	40% (31 of 77)	30 to 51%	51% (35 of 69)	39 to 62%

Notes to Table 8.1.

1. A nil null hypothesis is one of no effect, no difference, zero correlation etc.
2. If the author did not preface 'significant' with 'statistically', or follow it with a *p* value or test statistic, or otherwise differentiate statistical and substantive interpretations, the practice was recorded as 'ambiguous'.
3. For example, p=.003.
4. Use of asterisks (one, two, or three star significance) has been heavily criticized (e.g. Meehl 1978): it provides even less information than exact *p*-values, is usually insufficient for meta-analysis and has the potential to mislead researchers into thinking that an effect with two stars is more important than an effect with one.
5. For example, noting that the sample size was small.
6. Interpretation was any mention of: CI width; the possible theoretical importance of the upper or lower bound of the interval; overlap between two CIs; or the word 'precision' in relation to a CI.
7. Error bars included standard error and CI bars.

8.4 A Case Study of Statistical Reform in Ecology

Whilst it is too early to speculate broadly about what may lead to successful reform in ecology, I can offer a case study in how such change might occur. Professor Mark Burgman's Environmental Science laboratory at the University of Melbourne, Australia, provides this case. The students (roughly 12 PhD students) and post-doctoral fellows (roughly 5 post-doctoral fellows) in this lab are highly aware of the controversy over NHST. There is minimal reliance on NHST in their work and they instead rely on a range of techniques including: CIs, likelihood and AIC methods, Bayesian modelling, Bayesian networks, Bayesian credible intervals, Population Viability modelling and other sophisticated and still somewhat obscure techniques for decision making under severe uncertainty (for example, 'InfoGap', see Ben-Haim, 2001). They also include these ideas in teaching undergraduate and postgraduate courses in quantitative ecology, conservation biology and environmental risk assessment. On the rare occasions they do use NHST, they calculate statistical power with a specific a priori effect size.

But things weren't always this way. Mark Burgman's own undergraduate training was in Botany and Zoology where he was exposed to only classical NHST-based statistics, *excluding* statistical power and effect size. He relied solely on

statistical significance in his own PhD and much of his early career work (personal communication, July 2005). Burgman had studied under Professor F.J. Sokal, at the State University of New York at Stony Brook, USA. (Recall the Sokal was co-author of the first and perhaps still most widely read statistical text in ecology; Sokal and Rohlf is now its 3rd edition, 1995). Burgman remembers hearing about statistical power during this period—but it wasn't framed as particularly serious or important.

In 1990, when Burgman accepted a position at the University of Melbourne, he was introduced to Professor Michael Keough (Department of Zoology). Keough was the first to convince Burgman that type II errors were important. This in itself might not have been enough influence on Burgman, if it wasn't for another chance meeting the following year with Dr. Neil Thomason (Department of History and Philosophy of Science). Thomason convinced Burgman that the neglect of type II errors mattered and that misuse and misinterpretation of NHST were causing damage to science. It was understanding this final point—that damage was being done—that Burgman remembers as the turning point in his own practice.

Thomason and Burgman had a competition during one of their early lunches together. In which discipline—psychology or ecology—was the misunderstanding of NHST and neglect of statistical power worse, and where had the most damage been done? The consensus (I have asked both of them) is that Burgman eventually won with a story about the mortality rate of owls. A particular study Burgman had been sent to review had effectively zero statistical power: "management could have killed all the owls in field site and this study would still have concluded 'no significant difference'" (Mark Burgman, personal communication, September 2005).

Thomason also introduced Burgman to 'Statplay[32]', a software package designed to help overcome common statistical misunderstanding (designed by Geoff Cumming, Sue Finch & Neil Thomason). Burgman is now himself a prominent advocate of statistical reform in ecology (e.g., Burmgan, 2005; Burgman et al., 2000; Burgman & Possingham, 2000). Several of his advanced students, recent graduates and past and current post doctoral fellows could also be considered to fall into this category. For example—and this is far from an extensive list—they have published papers: drawing attention to statistical power problems in specialised areas (e.g., Carey & Keough, 2002;

[32] Statplay was never commercially released. However it is still used in teaching introductory statistics at the University of Melbourne, both in Environmental Science and by the Department of Mathematics and Statistics; for several years it was also used in the Psychology Department.

Wintle et al, 2004); applying and/or explicating information theoretic or Bayesian methods or Bayesian networks (e.g., Wintle et al, 2003); using modelling techniques, including Population Viability Assessments (e.g., Elith & Burgman, 2003). A complete list of publications can be found at: (http://www.botany.unimelb.edu.au/envisci/).

Few of them had heard of problems with typical NHST practice before coming to Burgman's lab. In an informal survey, current lab members told me when they first became aware of the problems we are all now familiar with. Some typical responses included:

> 'In 1998, when I started my PhD with Mark [Burgman]'
>
> 'A year ago when I started the PhD here. It was a general consensus I picked up in the lab, and probably from Mark [Burgman]'
>
> 'When I came here in mid-2003'

Only one person acknowledged having being alerted to problems earlier, and her response was heavily qualified:

> 'First heard this 6 years ago from the statistical consultant to staff and students at the Australian National University… but didn't get the full story on just how problematic it was then. I was only enlightened and apprised of all the details when I got to this lab.'

When I asked what published articles had influenced them or that they would recommend to others, several noted the contributions of David Anderson and Ken Burnham (specifically Anderson et al., 2000 and Burnham & Anderson, 2002), who primarily advocate an information theoretic model selection and model averaging approach. Another common reference was to Johnston's (1999) "The Insignificance of Statistical Significance". (Johnston and Anderson et al. were both published in the *Journal of Wildlife Management.*) However, despite there being virtual consensus over which articles have been most influential and helpful, the more pronounced influence on members of the environmental science lab is an internal force—first, their initial discussions with Burgman, and then later, conversations with each other.

Burgman may well do for ecology what Charles Poole (personal communication, quoted in Chapter Six) claims Rothman did for epidemiology. Like Rothman, Burgman sits on several editorial boards and is involved in many of the discipline's societies. The collaborative nature of work in this lab, and perhaps ecology more generally, also seems to play an important role in the spread of reform. For example, three of the PhD students in Burgman's lab work on different aspects of a

single industry funded project about climate change in the Alpine regions of Australia. If one makes a decision (as happened recently) to use Bayesian network analysis, it compels the others to at least include one comparable analysis of this kind.

8.5 Summary and Conclusion

Statistical reform has a short history in ecology, much of which was taken up with unhelpful debates over retrospective statistical power (using the obtained effect size). With this matter finally settled (and the practice deemed invalid) reform has moved to likelihood methods, notably information theoretic methods, and Bayesian approaches. Reform advocates have perhaps been more varied in ecology than in medicine or psychology. A few have recommended CIs, others the methods just mentioned, and there has also been a major focus on sophisticated and explicit modelling techniques.

There are promising signs of change in the discipline. Results of the journal survey presented in this chapter demonstrate a potential lessening of NHST in this discipline. However, reform in ecology is still in nascent stages and it is perhaps too early to speculate about successful interventions or motivations. If the changes surveyed in the conservation biology journals we sampled are real, representative of the discipline, and increasing, then such questions will soon become important in their own right.

9
CONFIDENCE INTERVALS AND STATISTICAL REFORM

Now people talk about 'precision' all the time, but what they mean by precision is whether the interval includes the null value or not… They think a result is 'imprecise' if the interval includes the null value, regardless of how wide it is. That is how strong the hypothesis testing stranglehold is: It is just so easy to see whether that null value is in the interval or not. It's a tremendously hard habit to break even for people who want to (Charles Poole, personal communication, September 2001).

As we saw in Chapter Two, CIs are sometimes advocated on the grounds that they communicate more information than *p* values and loose none. Statistical significance can be read from a CI, as the above quotation from Poole suggests (and somewhat laments). If the null value falls within the interval, the results are statistically non-significant at *p* equivalent to *C*, where *C* is the confidence level (e.g., 95%). But CIs offer extra information as well. First, they make effect sizes salient; a *p* value, on the other hand, provides no *direct* measure of effect. Second, CIs provide immediate information about the precision of study; this comes in the form of the width of the interval.

However, there is no guarantee that CIs will be a panacea to all statistical reporting woes: There are two main obstacles. First, as Poole pointed out, if researchers ignore the extra information contained in the interval—if they attend only to whether or not the null is captured—then CIs do not necessarily offer them more than NHST. Second, CIs, like *p* values before them, may be prone to misinterpretation. As Abelson (1997) warned, "Under the Law of the Diffusion of Idiocy, every foolish application of significance testing is sooner or later going to be translated into a corresponding foolish practice for confidence limits." (p. 130). At the heart of Abelson's claim is an empirical question.

Given that the high level of misinterpretation of NHST is the most common argument against continued reliance on it, it seems reasonable to expect whatever is proposed to be NHST's replacement be at least relatively free of such misinterpretation. But do CIs lead to fewer or less serious misinterpretations (or, the flip side, richer, more substantial interpretations) of research results than NHST? Surprisingly, little attention has been paid to this question. Advocates have often assumed that CIs will be relatively

intuitive and infrequently misunderstood. For example, Schmidt and Hunter (1997) claimed that CIs are easier and less frequently misinterpreted than *p* values, appealing only to teachers' anecdotal experience.

> Point estimates and their associated CIs are much easier for students and researchers to understand, and as a result, are much less frequently misinterpreted. Any teacher of statistics knows that it is much easier for students to understand point estimates and CIs than significance testing with its strangely inverted logic. This another plus for point estimates and CIs (p.56).

Elsewhere Schmidt also claimed that using CIs rather than *p* values would make research literature less confusing:

> …the use of point estimates of effect size and CIs in interpreting data in individual studies would have made our research literature far less confusing, far less apparently contradictory, and far more informative than those that have been produced by the dominant practice of reliance on significance testing (1996, p. 122)

Similarly, Hammond (1996) called CIs "simple and informative" (p.105). Quotations like these are not difficult to find in statistical reform literature. Some have gone so far as to suggest that even if CIs were misinterpreted, the misinterpretations may not have as serious consequences as those associated with *p* values.

> Although we cannot demonstrate it formally, we suspect that imperfectly understood confidence intervals are more useful and less dangerous than imperfectly understood p values and hypothesis tests. For example, it is surely prevalent that researchers interpret confidence intervals as if they were Bayesian credibility regions; to what extent does this lead to serious practical problems? (Hoenig & Heisey, 2001, p.23).

All of the above claims, and others like them, are made without supporting evidence, beyond the anecdote itself. In fact, very little is known about how researchers think about CIs, let alone how they might be misused or misinterpreted. For example, the inverse probability fallacy, as it relates to *p* values, has been accused of leading researchers to neglect statistical power and *a priori* sample size calculations (Oakes, 1986). How does this fallacy manifest, if at all, in the interpretation of a CI? As a Bayesian credible interval (as Hoenig & Heisey, 2001 suggest above)? What

implications would such a misconception have on the conclusions researchers draw from their research or the planning practices they engage in?

As I noted in Chapter Two, because CIs rely on the same sample information as NHST, and belong to the same philosophy of statistics (frequentist), some researchers may be tempted to think they are 'the same thing' as NHST. This dismisses a mass of evidence that different formats of equivalent information can profoundly affect our ability to complete conceptual algorithms and reason using the information. One important feature of more successful formats—those that allow us to best use quantitative information to reason—is that they have shorter *information menus* or fewer pieces of separated information (Gigerenzer & Hoffrage, 1995). CIs certainly have this feature: They combine information on effect size and precision, and can themselves be used to conduct significance tests if required (by determining whether the null value is in or out of the interval). Significance tests, on the other hand, require the separate reporting of a *p* value, an effect size and a statistical power calculation to provide equivalent information. The information in the latter format is therefore more fragmented, and undoubtedly more difficult to integrate. As a consequence of emphasising effect sizes and precision, CIs should facilitate meta-analytic thinking (Cumming & Finch, 2001).

The two studies reported in this chapter, along with the two reported in Chapter Ten, are preliminary efforts at establishing an evidence-based reform. They are empirical studies of students' understanding of CIs.

9.1 Do Confidence Intervals Help Avoid Established Misconceptions?

One particularly damaging misconception associated with NHST is that a statistically non-significant result is strong evidence for no effect (no difference, no impact etc). Published researchers often misinterpret statistical non-significance in this way, without any consideration of statistical power or the precision of the study (see surveys in Chapters Four and Eight). Because CIs make information about precision salient, particularly when they are presented graphically, we should expect that CIs would result in fewer misinterpretations of this kind than traditional NHST presentations.

9.1.1 Method

I surveyed 79 final-year Bachelor or Masters students from three separate environmental science classes—'Environmental Risk Assessment', 'Environmental Problem Solving' and 'Environmental Risk Assessment (Intensive)'—at the University of Melbourne[33]. All students had at least one prior semester of statistics, and were more than half way through a second quantitative course in risk assessment or environmental problem solving.

Students were randomly assigned two of four possible scenarios with fictional data and asked to answer some multiple choice questions. All scenarios reported statistically non-significant results, for studies with low statistical power (38-60%) and ecologically non-trivial observed effect sizes. By non-trivial I mean that in two scenarios the observed effect size was only slightly less than what was identified as a biologically important effect size and in the other two, slightly more than a biologically important effect size. All scenarios were simple research designs, either a single sample or two independent groups. The content of the four scenarios varied from soil and water contaminants with potential human health effects, to the population decline of popular and endangered flora and fauna. Every effort was made to match scenarios for perceived environmental importance, interest and accessibility. Scenarios were developed in consultation with two PhD ecologists (Dr. Sarah Bekessy and Professor Mark Burgman) to improve plausibility and symmetry.

In one version of each scenario, the results were presented as a *t* test. In these cases the *t* statistic was accompanied by: corresponding mean difference, standard deviation, degrees of freedom, *p* value and *a priori* statistical power calculation for a predetermined biologically important effect and a graphic of the mean difference. In the other version of each scenario, the results were presented graphically with CIs. In each individual survey, the scenario introductions were followed by either an NHST version of the results or a CI version. Below is a list of the four scenarios used.

> ***Scenario 1.*** "Toe-clipping is commonly used to mark frogs in population ecology studies because other methods of marking don't work on their skin. It is a valuable technique but there is some controversy over

[33] The University of Melbourne is one of the top three universities in Australia, and entry is extremely competitive. These were very bright students who had been taught by well respected academics. At the time of the survey, students were enrolled in one of the listed courses which were taught by Professor Mark Burgman.

whether it affects recapture rates and, therefore, frog survival. This study examined the decline in recapture rate of frogs that had toes clipped…"

Scenario 2. "Cadmium is a toxic heavy metal used, amongst other things, to make batteries. The cadmium level in stream near a battery factory has just been monitored…"

Scenario 3. "A new park land is being developed near an old gas works. A lot of soil has already been cleaned and replaced. Since the clean up, the concentration of petroleum has been surveyed…"

Scenario 4. "The monkey puzzle tree is a vulnerable species, endemic to South America. A study recently investigated whether two populations of monkey puzzle trees could be mixed for reforestation. If there are sufficient genetic differences between the two populations they should be kept separate; if not, they can be mixed. One important and common measure of genetic difference is the "root to shoot" ratio, which measures how drought tolerant the trees are…"

In each scenario, students were given explicit information of the null hypothesis.

Scenario 1. Zero decline in frog recapture rate.

Scenario 2. Normal background level of cadmium approximately = 1ppb.

Scenario 3. Average non-harmful level of petroleum = 2000mg/kg.

Scenario 4. Root to shoot ratio of 1.

In addition, students were provided with information about the size of a biologically important effect. In keeping with typical practice, the biologically important effect did not correspond to the nil hypotheses tested as nulls.

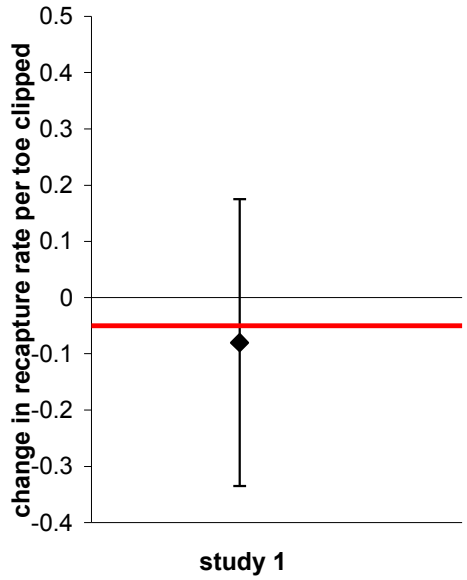
Scenario 1. Frog recapture rate decline of 10%.

Scenario 2. Environmental Protection Agency maximum acceptable level = 5ppb cadmium.

Scenario 3. 5000mk/kg of petroleum is dangerous to human health.

Scenario 4. Root to shoot ratio of 5 or above is believed to be of substantial genetic importance.

Confidence Interval Format



Toe-clipping is commonly used to mark frogs in population ecology studies because other methods of marking don't work on their skin. It is a valuable technique but there is some controversy over whether it affects recapture rates and, therefore, frog survival.

This study examined the decline in recapture rate of 60 frogs that had toes clipped.

In the figure above, the Y axis shows proportion change in recapture rate (negative values show proportion decline, positive values show increase). The black horizontal line crossing at 0 indicates no effect on recapture rate. The thicker, horizontal line crossing at -.05 indicates the minimum decline we understand to be ecologically unacceptable. If the true proportion decline exceeds .05, toe clipping is an unacceptable practice. The black diamond is the change in recapture rate for this sample; the error bar is a 95% confidence interval.

NHST Format

Toe-clipping is commonly used to mark frogs in population ecology studies because other methods of marking don't work on their skin. It is a valuable technique but there is some controversy over whether it affects recapture rates and, therefore, frog survival.

This study examined the decline in recapture rate of 60 frogs that had toes clipped. The minimum ecologically unacceptable decline in recapture rate is known to be .05. If the true proportion decline exceeds .05, toe clipping is an unacceptable practice.

The proportion decline in this sample was .08. This proportion (.08) is statistically *not* significantly different from zero (one sided t test = 1.1, p =.27; df = 59). The *a priori* statistical power of this test, to detect a decline of .05, was 40%.

Figure 9.1. The 'toe-clipping' scenario in two formats—CI graphic and NHST text. In both the effect is biologically important and statistical power (or precision) low.

Figure 9.1 shows examples of one scenario (toe-clipping of frogs) in both formats. In NHST scenarios, students were told the *a priori* statistical power of detecting these biologically important effects. In CI scenarios, both the null and the

biologically important values were marked on the error bar graphic. Students were given the following question and asked to answer by circling one of the five statements:

> In response to this information, the researcher who conducted this study should conclude that:
>
> - There is strong evidence in support of an important effect.
> - There is moderate evidence in support of an important effect.
> - The evidence is equivocal.
> - There is moderate evidence of no effect.
> - There is strong evidence of no effect.

The exact wording of these statements changed with each scenario depending on what the particular 'effect' was. For example, in Scenario 1 the final option read: "There is strong evidence that toe clipping does not cause unacceptable decline". In Scenario 2, it read: "There is strong evidence that the factory has not breached EPA [Environmental Protection Agency] standards." (Underlining in original.) The order of responses was reversed in *Scenario 3*.

Classification of Responses

I classified as a misconception statements of moderate or strong evidence for the null hypothesis. In all scenarios the statistical power of the study was low (power between 38% and 60%) and effect sizes were non-trivial in comparison to biologically important effects. Therefore, accepting or 'failing to reject' the null was an uncontroversial error, and entailed interpreting statistical non-significance as 'no effect'. More appropriate responses either: noted the lack of power or precision and deemed the evidence equivocal; or attended to the large effect sizes and suggested that the evidence favoured the alternative hypothesis.

9.1.2 Results and Discussion

Surprisingly, 61% (48 of 79, 95% CI: 50 to 71%) of students did *not* demonstrate the misconception that statistical non-significance means 'no effect' when given results presented in the NHST format. This in itself impressive. Previous research has found this misconception to be far more widespread (Haller & Krauss, 2002; Oakes, 1986). However, it is important to note that these students had, throughout the semester, received several warnings of the misconception and formal

instruction regarding statistical power analysis. Also, statistical power was clearly stated in all scenarios, and the biologically important effect size was stated. This is far from typical practice. For example, in the survey of conservation biology journals reported in the Chapter Eight, only 8% of articles reporting statistically non-significant results in 2005 reported a statistical power calculation, yet almost half of the articles survived interpreted statistical non-significance as evidence for no effect. Given that the result presentation in these scenarios was much more complete than a typical journal article, it should perhaps be alarming that still 39% (31 of 79; 95% CI: 29 to 50%) *did* demonstrate the misconception.

Of the students who demonstrated the misconception in the NHST scenarios, an overwhelming majority (87%, 27 of 31; 95% CI: 71 to 95%) gave correct answers to the CI scenarios (presentation order was counter-balanced). Only 4 of those 31 students with gave the same answer regardless of presentation; the rest showed improved interpretations when presented with the same data in the CI format. Almost a third (32%, 10 of 31; 95% CI: 19 to 50%) of those who had demonstrated the misconception with NHST moved 1 point on the 5 point likert scale *away* from statements of accepting the null when given a CI; roughly half (52%, 16 of 31; 95% CI: 35 to 68%) moved two points and 3% (1 of 31; 95% CI: 0 to 16%) moved three points. This amounts to an average shift of 1.67 points on a 5 point scale.

So far, this is an undeniably impressive result in favour of CIs. But what of students who did not demonstrate the misconception to start with? Were they perhaps led astray by the CI format? Recall that 61% of students did not demonstrate the misconception in the NHST scenario. Of those, 17% (8 of 48) *did* demonstrate the misconception when given the CI scenario. This 'reverse' effect is obviously undesirable. Ideally, there would be no shift in this direction. However, there was considerable variation between scenarios, despite our best efforts to match them. Figures 9.2 and 9.3 show the distribution of responses in each scenario for NHST and CI formats. The advantage of CIs was clearest in the 'frogs', 'cadmium' and 'petroleum' scenarios. It was less clear in 'monkey puzzle' scenario—in fact CIs had a small negative impact in this scenario. Table 9.1 shows the percentage of students who demonstrated the misconception for each scenario and format.

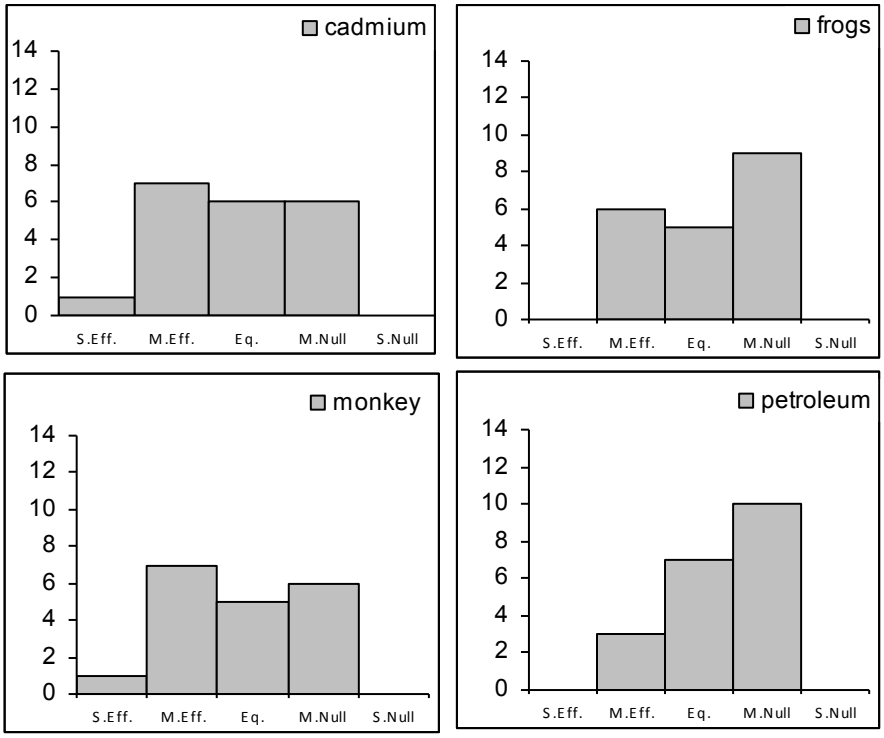


Figure 9.2. Frequency of five response types (Left to Right: Strong Support for Effect, Moderate Support for Effect, Equivocal, Moderate Support for Null, Strong Support for Null) for the four scenarios when results were presented in NHST format. (Number of respondents in each scenario is not equal: Cadmium n= 20; Frogs n=20; Monkey n=19; Petroleum n=20).

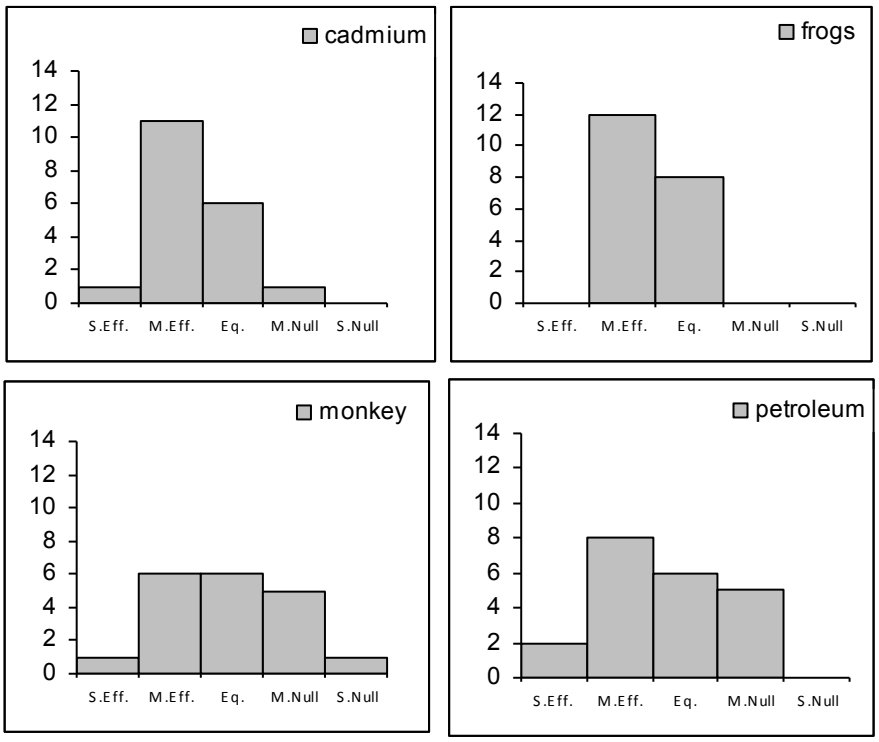


Figure 9.3. Frequency of five response types (Left to Right: as above) for the four scenarios when results were presented in CI format. (Number of respondents in each scenario is not equal: Cadmium n= 21; Frogs n=20; Monkey n=19; Petroleum n=21).

Table 9.1.
Percentage of students who agreed that results moderately or strongly supported the null hypothesis (i.e., that demonstrated the misconception that statistical non-significance equals 'no effect') for each scenario and format.

	NHST % (n)	CI % (n)	Improvement[a] with CI (NHST-CI)	95%CI[b] for improvement
Cadmium	30% (6 of 20)	5% (1 of 21)	25%	3 to 50%
Frog	45% (9 of 20)	0% (0 of 21)	45%	23 to 67%
Monkey Puzzle	32% (6 of 19)	37% (7 of 19)	-5%	-35 to 25%
Petroleum	50% (10 of 20)	24% (5 of 21)	26%	-2 to 55%

[a]Negative values reflect more misconceptions in CI format of the scenario than in the NHST format.

[b]CIs calculated according to method recommended by Newcombe & Altman (2001).

9.1.3 Conclusion

At first glance the results show promise for CIs, even if a little less than expected. Given (a) variability across scenarios and (b) that NHST used was best practice (e.g., *a priori* power, biologically important effect size clearly stated) these results should be interpreted as encouraging. This misconception that statistical non-significance means no effect is not entirely absent when results are presented in CI format (as we might first have expected), but it is less frequent.

9.2 A Replication

The survey described in part one was partially replicated as a fully within subjects design in the hope of eliminating variability between scenarios that may have masked the positive effect of CIs. The sample was a class of 55 second year ecology students at the University of Melbourne. Arguably, these students were on average less statistically sophisticated that the students in the previous sample. They had varying levels of statistical experience, but all had at least been exposed to both CIs and NHST in the previous semester.

9.2.1 Method

Each student was given a single research scenario and presented with both CI and NHST presentations of the results. This replication was designed to eliminate any confounding effects of scenario content, including varying effect sizes, levels of statistical power and subjects of study. The scenario was introduced as follows:

> There are concerns about the air quality in a freeway tunnel. This study monitored the concentration of carbon monoxide (CO) during peak hour traffic over two weeks, taking a total of 35 samples. Normal background levels of carbon monoxide are between 10-200 parts per million (ppm). One hour exposure time to CO levels of 250ppm can lead to 5% carboxylated hemoglobin in the blood. Any level above this is abnormal and unsafe. If the true level of CO concentration in the tunnel exceeds 250ppm, the tunnel will be closed and a surface road built. However, the surface road proposal has problems of its own, including the fact that threatened species inhabit an area near the surface site. First consider Presentation A. Please answer the question following Presentation A and then move on to Presentation B.

Again, the scenario was statistically non-significant with low statistical power and an effect size close to the biologically important cut off (i.e., observed effect 230ppb). In addition, this scenario also included an incentive against overly precautionary answers—the biological cost of the tunnel alternative.

Question one asked whether the results provided strong or moderate evidence for an unsafe circumstance (the alternative hypothesis), strong or moderate evidence for a safe circumstance (the null) or whether the evidence was equivocal. This question was the same format as described for part one of this study. Question two asked students directly whether they thought the NHST format was misleading:

> Think about the two presentations you have just seen, A and B. Presentation A [NHST] has been criticised on the following grounds: 'It is likely to mislead people into thinking there is no difference between the observed value and the ordinary background level, when in fact the 95% CI in the other presentation shows that a true CO level of up to 270ppb would be consistent with this data.' How likely is that people would read presentation A as providing evidence for 'no difference'?

Students answered question two on a 5 point likert scale from ‘very likely’ to ‘very unlikely’.

Classification of Responses

As for the part one, I classified as a misconception responses that suggested results provided moderate or strong evidence for the null hypothesis. Because statistical power of the study was low and the effect size was non-trivial in comparison to biologically important effects, accepting the null was an uncontroversial error.

9.2.2 Results and Discussion

Interpretation

When asked to interpret results presented in NHST format, 44% (24 of 55; 95% CI: 31 to 57%) of students misinterpreted statistically non-significant results—from a low powered study with a non-trival effect size—as evidence for the null hypothesis. Less than half as many (18%, 10 of 55; 95% CI: 10 to 30%) made this mistake in the CI condition.

In part one there was a reversal effect. That is, of the students who did *not* demonstrate the misconception in the NHST condition, some (17% in part one) did demonstrate it in the CI condition. In part two, which eliminated confounding of certain scenarios, the reversal effect was much less pronounced. Only 6% (2 of 31) of those that did *not* display the misconception in the NHST condition did display it in the CI condition, suggesting that the variability amongst the four scenarios in part one may indeed have been responsible for disguising the positive effects of CIs.

The average shift away from misconceptions in NHST versus CI conditions was 1 point on the 5 point scale—perhaps not quite as large an effect as we might have anticipated. However, this shift may be conflated by a learning effect. Students who saw the CI first gave the correct answer on the *p* value presentation more often than students who saw *p* values first. There was no corresponding beneficial transfer of seeing the *p* values before CIs. (Numbers here are too small to analyse in any formal way.)

Rating of NHST Format

Four students failed to answer question two, so percentages here are calculated out of 51 rather than 55. Table 9.2 shows percentages of students who answered that results presented in NHST format were or were not likely to be misleading. Almost two thirds (65%) of students answered that NHST was either very likely or likely to mislead; 16% disagreed, and 20% indicated they didn't know. Unfortunately, I did not ask a corresponding version of this question for CIs.

Table 9.2.
Percentage of students who agreed NHST format was misleading.

	%	95% CI[1]
NHST very likely to mislead	12 (6 of 51)	5 to 23%
NHST likely to mislead	53 (27 of 51)	40 to 66%
Don't know	20 (10 of 51)	11 to 33%
NHST unlikely to mislead	12 (6 of 51)	5 to 23%
NHST very unlikely to mislead	4 (2 of 51)	1 to 13%

[1]CIs calculated according to method recommended by Newcombe & Altman (2001).

9.2.3 Conclusion

This within-groups replication of the survey in part one provided even stronger evidence that CIs have a cognitive advantage over *p* values. CIs substantially help alleviate the misinterpretation of statistically non-significant results as evidence for the null hypothesis (despite information on low statistical power). With variation between scenarios eliminated, the positive effects of CIs were obvious, producing an average 1 point shift on the 5 point scale. Problematic 'reverse' effects were dramatically reduced so that they only occurred in only 2 of 31 participants.

10
DO CONFIDENCE INTERVALS HAVE MISCONCEPTIONS OF THEIR OWN?

As the observed mean difference…lies within the 95% CI we can conclude that this mean is a plausible value and that there is a difference (Student interpretation of CI, in study one, Chapter Ten).

To fulfil the role of 'NHST substitute' CIs need to be free not only of the worst misconceptions associated with NHST, but also of any other previously unknown misconceptions. In short, they ought not bring serious misconceptions of their own, above and beyond those of NHST. At the very least, if there are misconceptions particular to CIs, we want them to have less serious consequences or be easier to overcome than those associated with NHST.

The two studies presented in this chapter were designed to investigate students' understanding of CIs, and identify any potential misconceptions associated with their interpretation. The first study was an exploratory, open-ended survey comparing interpretations of NHST and CI presentations of results. The second was a somewhat more structured survey of CI understanding only, designed to further investigate some of the typical CI interpretations uncovered in the exploratory survey.

10.1 Exploring Students' Interpretations of Research Results

This study examined students' *interpretation* of results presented as either *p* values or CIs: It did not address the *calculation* of these values. Responses were open-ended to investigate differences in 'type' of response, as well as correctness, between groups given different presentations. The survey was specifically designed to investigate reformers' claims that CIs are less susceptible to misinterpretations and misunderstandings than NHST. The study also partially replicated Tversky and Kahneman (1971), by asking students to estimate the number of subjects needed to replicate the experiment outlined in the survey scenario. Students were also asked to make predictions regarding the replication study (e.g., what effect size they expected).

10.1.1 Method

I surveyed 177 first year psychology students at the end of a year long introductory statistics course at La Trobe University. The statistics course was introductory, but students had been exposed to some criticisms of NHST and the course had been heavily focused on CIs[34]. We could therefore expect that students in this course had more than average experience with CIs. Two independent groups were given fictional experimental data and asked to interpret resuls. One group was given results presented in CIs (n=95) and the other, the same results presented with p values (n=82). The fictional scenario was designed to be typical of experimental results in psychology.

> A researcher is interested in the effect of marijuana on reaction time. She conducts an experiment on 40 undergraduate students. She measures reaction time by time-taken-to-brake in a driving simulation task. Each student takes the simulation task twice—before and directly after smoking marijuana. The time-taken-to-brake is the reaction time measure, and the difference after-minus-before is calculated. The results show that the mean difference in reaction time (after-before) was 4.3 seconds (s=10.0 seconds).

In the CI condition students' were presented with the following analysis:

> The researcher finds a 95% CI for the mean difference in reaction time is: 1.1 to 7.5 seconds.

In the NHST condition, the equivalent results were presented as:

> The researcher conducted a hypothesis test (two tailed t test, alpha=.05) to see whether marijuana has a statistically significant effect on driving time. Her null hypothesis is that the mean difference in time-taken-to-brake is 0. The value of t is 2.7 and the test revealed p<.05.

As the results presentation had been designed to reflect typical reporting practice in psychology journals, statistical power calculations were not reported for the NHST scenario. (I realise this is somewhat controversial, but the decision to mimic journal reporting practices, rather than 'best practice', was made after considerable thought). A short series of questions related to interpretation of the findings followed. The first question was open-ended and general "What conclusions would you draw from these

[34] The course was taught by Dr. Geoff Cumming.

results?" This broad question was designed to reveal differences in both correctness and type of response, for example, whether responses to CI presentation were more often correct, and also whether they were longer and more substantial in content than NHST answers (or vice versa). The second question partially replicated Tversky and Kahneman's (1971):

> As the researchers' results will be used to inform new legislation about driving under the influence of marijuana, she intends to run another study to check her outcomes. How many subjects should she use in the second study? What do you expect the second study to find (e.g., what might the mean difference in the second study be)?

This question was designed to assess students' understanding of sampling variability and how it might be affected by the results format.

10.1.2 Student Responses: Identifying 'Effects' and 'Differences'

Coding criteria was developed after reading through the students' responses—it was therefore descriptive of their responses, rather than predetermined. Coding categories are described below and some sample responses are offered to help define categories. Students' responses often fitted more than one category, so the sum of responses across categories does not equal the number of students. Proportions of responses (with 95% CIs) in each of the categories outlined below are given below in text and in Table 10.1.

Salience of the Effect Size

To measure how salient the effect size was to students interpreting results, I noted any mention of the mean difference in their responses. Overall, very few students (7%, 13 of 177; 95% CI: 4 to 12%) explicitly mentioned the mean difference. In the NHST group, only 1 student of 82 did this; in the CI group, reference to the mean was more common than the NHST group although still surprisingly infrequent at only 13% (12 of 95; 95% CI: 7 to 21%).

Statements of 'Effect' Versus Statements of 'Difference'

For this category, I attempted to differentiate substantial statements of effect, from statements that simply acknowledged a difference between groups. Statements

coded as 'effect' focused on the direction or size of the effect and, in general, directly implied the experimental treatment was responsible for changes in the dependent variable (i.e., identified a casual relationship). Example statements of 'effect' are below.

"Reaction time is increased by marijuana"

"Marijuana slows driving time"

Twice as many CI group students made statements of effect as NHST group students: 22% (21 of 95; 95%CI: 15 to 31%) in the CI group; 10% (8 of 82; 95% CI: 5 to 18%) in the NHST group. When a statement implied a casual effect, but failed to specify the direction of the effect it was coded as a 'non-specific statement of effect': for example, "Marijuana affects reaction time". Similar proportions of these non-specific statements appeared in each group: 22% (21 of 95; 95% CI: 15 to 31%) in the CI group and 21% (17 of 82; 95% CI: 13 to 31%) in NHST group.

Statements of difference, rather than the effect, did *not* indicate the casual relationship between experimental intervention (marijuana) and the dependent variable (reaction time) or the task it measured (braking time when driving). Responses in this category never mentioned the direction or magnitude of the effect.

"Results suggest there is a difference in reaction time"

"There is a difference in driving times"

In the CI group, just 4% (4 of 95; 95% CI: 2 to 10%) gave responses of this kind, compared to 12% (10 of 82; 95% CI: 7 to 21%) in the NHST group.

Statements which referred to an 'effect' and particularly the direction of the effect were considered more sophisticated than mere statements of difference. On this account the CI group performed slightly better than the NHST group making 12% more (95% CI: 1 to 23%) full statements of effect, roughly the same non-specific statements of effect (1% more; 95% CI=0 to 11%) and fewer simple statements of difference (8% less; 95% CI=0 to 17%). Of course, these proportion differences between the CI and NHST groups are small and the CIs are reasonably wide, so conclusions remain tentative.

Statistical Significance and the existence of an Effect

Overall, 12% (21 of 177; 95% CI: 8 to 18%) of students misinterpreted the results as evidence of no effect. 'No effect' was considered a misinterpretation as the

scenario results were clearly statistically significant and effect size substantial. Student responses of this kind were of the following types:

"Has no effect"

"Results show the use of marijuana has no effect on driving time"

"There is no significant difference"

More students in the NHST group made this mistake: 20% (16 of 82; 95% CI: 12 to 29%) of the NHST group compared with just 5% (5 of 95; 95% CI: 2 to 12%) of the CI group. I can only assume these students attempting to employ a rule of statistical significance (i.e. statistical significance=$p<.05$ or 'if zero is not in interval') which they had misremembered, and essential answered backwards.

Unable to Interpret

Of the 177 students in the survey, 5% (9 of 177; 95% CI: 3 to 9%) gave answers that suggested they were unable (or unwilling) to interpret results. These responses were either blank or answered with "I don't know". Such responses were slightly more common in the NHST group (7%, 8 of 82; 95% CI: 5 to 18%) than in the CI group (3%, 3 of 95; 95% CI: 1 to 9%).

Table 10.1.
Percentage of students who gave responses of the listed category types.

	% of NHST group (*n* of 82)	95% CI	% of CI group (*n* of 95)	95% CI
Mentioned effect size	1% (1)	0 to 6%	13% (12)	7 to 21%
Statement of effect	10% (8)	5 to 19%	22% (21)	15 to 31%
Non-specific statement of effect	21% (17)	13 to 31%	22% (21)	15 to 31%
Statement of difference only	12% (10)	7 to 21%	4% (4)	2 to 10%
Non-sig means 'no effect'	20% (16)	12 to 29%	5% (5)	2 to 12%
Unable to interpret	7% (6)	3 to 15%	3% (3)	1 to 9%

Using Rules

Many students relied on explicit rules to answer question one and to draw conclusions from the results presented. By rules I mean a statement regarding: a)

statistical significance; b) rejection of the null hypothesis; or c) the presence of absence of an effect, linked (by 'therefore', 'because', 'due to') to a statement regarding the *p* value, *t* value or bounds of the CI. In coding responses, I attempted distinguished between plausible and implausible rules. Figure 9.4 provides some examples of each category from students' responses.

Those in the NHST group worked more with rules (both plausible and implausible) than those in the CI group: 35% (29 of 82) of students in the NHST group used rules, compared with just 5% (5 of 95) of students in the CI group. In the NHST group, rules were plausible in 59% of cases (17 of 29), and implausible the rest of the time (41%, 12 of 29). Of the 5 students in the CI group that relied on rules, 2 used plausible rules and 3 implausible rules. The examples of implausible rules used with CIs given above provide an introduction to the sort of misinterpretations we found of CIs. The following section deals specifically with the CI group, and specific errors of interpretation they made.

10.1.3 More About Student Responses: CIs as Descriptive Statistics?

A pattern in responses of the CI group suggested widespread misunderstanding of what it is a CI estimates, or indeed, that a CI is an estimate—an inferential statistic—at all. Roughly half (47%; 45 of 95) of students in the CI group made an attempt define, describe, or make some substantive statement about how a CI is interpreted. By this I simply mean went beyond simple or vague statements such as 'there is an effect' or 'reaction time is slower'. Of those that interpreted the interval directly, 20% (9 of 45) of responses were plausible. Some examples of plausible interpretations of a CI:

"95% confident that the population mean would lie in this interval"

"95% confidence the population mean is between 1.1 and 7.5 seconds"

Obviously, I am not concerned here with a strict or technically correct frequentist definition but, more fundamentally, an interpretation that acknowledges the inferential nature of a CI. The remaining 80% (36 of 45) of interpretations provided by students failed to reach even that minimal standard: 56% clearly thought of the CI as a descriptive statistic and 24% of responses were ambiguous or incoherent. These erroneous interpretations (and the frequency of their expression) are listed in Table 10.2.

Plausible Rules:

NHST group

"Marijuana effects reaction time because the H_0 had to be rejected because $p<.05$"

"The *t* test provides evidence that H_0 cannot be supported as it has an extreme value"

"Results are statistically significant because the score revealed $p<.05$"

CI group

"Since 0 is not within the CI we can be 95% confidence that the marijuana has had an effect on reaction time"

Implausible Rules:

NHST group

"Reject H_0 as the value of *t* is greater than zero"

"There is no effect because $p<.05$"

CI group

"4.3 lies within the interval 1.1-7.5, therefore it is acceptable"

"95% CI contains the mean, therefore 95% confidence that the sample represents the population"

Figure 10.1. Examples of plausible and implausible rules in the NHST group and the CI group.

Table 10.2.
Percentage of student responses which included each of the listed erroneous definitions of a confidence interval.

	% (*n* of 45)	95% CI
CI estimates sample mean	18% (8)	1 to 31%
CI is range of individual scores	9% (4)	4 to 21%
CI is truncated range of individual scores	29% (13)	18 to 43%
Ambiguous	24% (11)	14 to 39%

CIs 'Estimate' the Sample Mean.

This category of response was very surprising, and might be difficult to imagine. Yet 18% of students clearly indicated that they believed the CI to be an estimate of the sample mean. For example:

> "As observed mean difference in RT lies within the 95% CI we can conclude that this mean is a plausible value and that there is a difference in time take to brake"
>
> "95% confident that the sample mean will fall within the range 1.1-7.5"
>
> "4.3 seconds is within the interval and therefore plausible"
>
> "Mean difference is 4.3, which is within the interval 1.1 to 7.5"

One student articulated the precise confusion:

> "95% sure that the sample (population?) mean is between 1.1 and 7.5 seconds"

CIs provide the Range of Individual Scores

Further on the theme of the CI being a descriptive, rather than an inferential, statistic was the category of response that equated the CI with the range of the raw data, or individual scores. Just 9% of students did this in a straightforward way, but several more (an additional 29%) defined the CI as a truncated range (i.e., 95% of range). Straightforward examples of the 'full range' type:

> "Reaction time ranges from 1.1 (not largely affected by marijuana) to 7.5 seconds (strongly affected)"
>
> "Average difference of about 4 seconds, with range from 1.1 to 7.5 seconds"

The following is an example of an interesting deviation on the 'CI is full range' type response:

> "The lowest RT occurs before marijuana was take (1.1 to 4.3); it takes longer (4.3 to 7.5) to react after marijuana"

Finally, the more common 'truncated range' response:

> "Participants are likely to have reaction times within this interval 95% of the time"
>
> "Marijuana slows RT by 4.3 seconds on average, differences larger than 7.5 seconds would only occur in 5% of cases"
>
> "95% of participants took between 1.1 and 7.5 seconds to brake after having marijuana"
>
> "95% of responses lay between 1.1 and 7.5"

Altogether some 38% of responses described the CI as the range of individual scores.

Ambiguous Responses

Of all attempts to describe or to directly interpret the CI, almost a quarter (24%, 11 of 45) were ambiguous or incoherent. They clearly failed to identify a CI as an inferential statistic, but they did not necessarily categorise it as descriptive either. These responses were quite varied, and had no obvious pattern. Here is one example: "Reaction time is reduced and interval appears smaller and slow."

10.1.4 Still More About Student Responses: Ideas About Replication

This section of the survey was modelled on Tversky and Kahneman's (1971) question about the sample size needed to run replication study (with similar results). The participants in this introductory class performed a lot better than Tversky and Kahneman's sample. In Tversky and Kahneman's study the median recommendation was for half the original number of subjects to be run in the replication. By contrast, 175 of the 177 students in the current survey recommended that the researcher run at least the same number of subjects or more. (The other 2 participants did not answer.) It is the second half of question two—"What do you expect the second study to find? (e.g. what might the mean difference in the second study be?)"—that I will discuss here in detail.

There were 21 uncodable responses in this section, reducing the overall *n* to 156 (*n*=72 for NHST group; *n*=84 for CI group). Over half the students (53%, 82 of 156) answered "results will be same as the first study" or "similar to the first study".

Accuracy

In the CI group, 30% (25 of 84) said they expected an increase in the accuracy of the results of the replication; 21% (15 of 72) in the NHST group also expected this. Sometimes, these responses were relatively vague, for example "the accuracy will be increased". Others, however, were quite detailed. They explained that the expected increased accuracy would be due to "a narrower CI", "reduced standard error" or "increased power". These are plausible answers, particularly given that the overwhelming majority of students expected the replication study to have a 'similar or larger' sample size.

Effect size

There was a reasonably widespread belief that the effect size would increase with replication. This was particularly pronounced in the NHST group where 21% (15 of 72) expected "a bigger effect", "an increase in mean difference" or "slower reaction times". In the CI group, 10% (8 of 84) expected an increase in effect size. Curiously, an additional 10% of the CI group and 7% (5 of 72) of the NHST group expected a *decrease* in effect size. Unfortunately, most responses were too vague to determine whether the students were in error (increasing sample size should not necessarily increase or decrease the effect size) or whether they were appropriately attempting to express recognition of sampling variability.

10.1.5 Conclusion

The most striking finding of this survey was that so many students failed to identify the CI as an inferential statistic, instead interpreting it as descriptive statistic—either bizarrely interpreting it as an estimate of the *sample* mean, or as something equivalent to the range (or truncated) range of individual scores. Despite this, students presented with CIs, rather than NHST results, made more substantial interpretations of effect, as opposed to mere difference statements, and drawer fewer incorrect conclusions about the overall results of the experiment.

10.2 A Semi-Structured Replication

The second study focused specifically on the interpretation of CI and the misconceptions identified in students' open-ended responses in question one of part one. In this study, I investigated two aspects of CI understanding. First, *definitional* understanding of CIs, that is, understanding the definition of a CI, that it is inferential and what it estimates. Second, *relational* understanding of CIs, that is, how the various determinants of a CI affect each other (e.g., that a CI gets wider as the confidence level increases).

10.2.1 Method

I surveyed 180 undergraduate students who had completed between one and four semesters of statistics. This sample was admittedly mixed but all students had at least been taught CIs at an introductory level. There was no overlap with students who completed the survey in part one. The survey in part two was, unlike part one, compiled of fixed response questions. Students were required to either circle correct answers from a multiple choice, or tick items of a checklist. Their chosen responses were categorised as correct (plausible) interpretations and misconceptions.

The scenario used was virtually identical to that in part one, only the mean difference (after-before) was adjusted slightly, to be more realistic (after further consideration, the mean difference in part one were deemed too large to be plausible). Here all students (n =180) were given CIs in the results, and asked various questions about how to interpret them. Some extra information, on the range of the data, was also added. I expected presenting the range of the data would help deter students from interpreting the CI as though it were itself the range. The scenario results were presented to students in the following format:

> The mean difference in reaction time (after-before) was 2.2 seconds
> Standard deviation = 5 seconds. Standard error = .79.
> The minimum value was .04 seconds; the maximum value 7 seconds.
> The 95% confidence interval was .5 to 3.9 seconds.

10.2.2 Results

When choosing from a list of items the CI might describe, less than a quarter of students (22%, 40 of 180) correctly identified the CI as a "range of plausible values for the population mean." The percentage of students agreeing with selected responses is shown in Table 10.3.

Table 10.3.
Percentage of students selecting each of the listed confidence interval definitions from a multiple choice list.

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

The percentage of students agreeing with incorrect descriptions of CIs further increased when the question format changed. For example, when choosing from the list above 38% of students agreed that the CI was a "range of plausible values for the sample mean". However, when asked: "The 95% confidence interval has a ___% chance of capturing the *sample* mean"[35], an overwhelming 84% (151 of 180) of students answered "95%". Only 16% (29 of 180) gave the correct answer of 100%. Even those that correctly selected 'population mean' from the original list performed poorly on this question: 66% (19 of 29) of these students also answered "95%". Clearly, students' understanding is fragile, as it is easily disrupted by changes in format. Only 6% (10 of 180) students got both questions correct.

Confusion over the sample and population means may directly impact students' understanding of CIs. One possible impact of this confusion is the misconception that a CI is a range of individual scores. As I discussed, this misconception arose unsolicited in part one and many students in part two confirmed they too held the incorrect belief. Whilst only 8% of students chose 'range' as a CI description (from the list in Table 10.3), a further 11% selected 'truncated range'. When presented with the following true or false question, "The 95% confidence interval covers 95% of the range of individual scores", almost two thirds (65%, 117 of 180) of students gave the wrong answer (i.e., "true").

The misconceptions identified above can be collectively identified as 'definitional misconceptions'. The second category of misconceptions I discuss is 'relational misconceptions'. These are misconceptions about the way different

[35] The word sample was clearly italicized in the survey.

components of a CI relate to each other, for example, how the confidence level affects the CI width. I asked students the following question:

> Using a much larger sample size is likely to [increase / decrease / have little effect on / unsure] the width of the confidence interval.

The correct answer is that the width should 'decrease'; only 16% (28 of 180) of students gave this answer. Over a third (36%, 64 of 180) were 'unsure' of the relationship, 29% (52 of 180) thought sample size had 'little effect on' the width and 20% (36 of 180) answered that they expected width to 'increase' with sample size. This indicates widespread confusion about how the components of a CI relate to each. This confusion was more pronounced when students were asked to relate the CI level and CI width. About three quarters (73%, 131 of 180) of students agreed that: "A 90% confidence interval for the same data would be wider than the 95% confidence interval". Of course, CI width decreases with the confidence level, and agreeing with the statement is an error. These results are summarised in table 10.4.

Table 10.4.
Percentage of students agreeing with relational statements about confidence intervals

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

[a]correct response; all others represent relational misconceptions

There is some evidence that students' may also have 'relational' misconceptions about effect size, although these appear to be less widespread than relational misconceptions about CIs. For example, 16% (29 of 180) students answered that "using a much larger sample is likely to [increase] the size of the difference between before and after". Another 11% (20 of 180) thought the size of the difference would *decrease*. The most plausible answer, which a majority (62%, 111 of 180) of students gave, was that increasing the sample size would have little effect on the size of the difference. The remaining 11% (20 of 180) answered 'unsure'. Most students demonstrated an

understanding of the relationship between sample size and statistical power. Table 10.5 shows percentage of students agreeing with various relational statements about effect size, statistical power and sample size.

Table 10.5.
Percentage of students agreeing with relational statements about effect sizes and statistical power.

	% (*n* of 180)	95% CI
Increasing *n* improves statistical power[a]	86% (155)	80 to 90%
Increasing *n* increases effect size	16% (29)	12 to 22%
Increasing *n* decreases effect size	11% (20)	7 to 17%
Increasing *n* doesn't necessarily impact on effect size[a]	62% (111)	54 to 69%

[a]correct responses; others represent relational misconceptions

10.3 Summary

The study in part two was designed to further investigate misconceptions about CIs that were uncovered in students' responses to the survey in part one. In particular, its purpose was to investigate the apparent false belief that CIs are merely descriptive statistics, offering information about the sample mean and the range (or truncated range) of the raw data. Students were also asked questions about how various aspects of a CI relate to each other.

In part one, 18% of students (who provided answers directly addressing the CI) spontaneously generated the misconception that a CI estimates the *sample* mean and 38% the misconception that it estimates the range, or truncated range. In part two, students demonstrated the same misconceptions. However, their susceptibility to these misconceptions was dramatically affected by question format, suggesting that their understanding is fragile. For example, just over a third (38%) of students chose "plausible values for the sample mean" as a description of a CI from a list (that included the more correct "plausible values for the population mean"). Yet, in a separate question, a vast majority (84%) claimed a 95% CI had only a 95% chance of capturing the *sample* mean. The percentage of students claiming the CI represented the range, or truncated range, of the raw data varied similarly. When choosing from a list just 19% chose one of the range options (8% range, 11% truncated range), yet 65% agreed with a

later statement that the CI covered 95% of the range of individual scores. Together results from parts one and two provide strong evidence that misconceptions about the inferential nature of CIs are widespread, and carry implications for statistical education, and reform more generally.

Students displayed a limited understanding of how sample size and confidence level relate to CI width. Only 16% realised that, all other things being equal, CI width would decrease (that is, the CI would be narrower) with sample size, and 73% agreed with a false statement that, for the same data, a 90% would be wider than a 95% CI. CI width is an important guide to the precision of results, and understanding how width is influenced by other aspects of the data is important to a full interpretation of research results using CIs. It was therefore disappointing to see such widespread misconceptions, especially from students whose curricula had included above average attention to CIs.

Yet despite misconceptions, CIs may still lead to richer, more substantial interpretations of research findings than NHST. The current results suggest that for the benefits of CIs (e.g., extra information regarding precision, greater potential for meta-analytic thinking) to be fully realised, statistics education needs to focus on guiding students through the relational properties of CIs, and dealing with the confusion between descriptive and inferential statistics.

10.4 Further Misconceptions about Confidence Intervals

Several of the misconceptions identified in the student samples discussed above are fundamental, and possibly would not persist past an undergraduate level (although I have no evidence to support this). However, there are other specific misconceptions about CIs that are widespread even amongst leading researchers and published authors. Below is a summary of a relatively new research program aimed at identifying and understanding the origins of such misconceptions.

10.4.1 The Overlap Misconception

Many researchers falsely believe that for two independent group means to be statistically significantly different the 95% CIs around those means must not overlap (or can ‘just touch’). In fact, 95% CIs can overlap by roughly a quarter and the difference

still be significant at $p<.05$ (Cumming & Finch, 2005). We (Belia, Fidler, Williams & Cumming, in press) demonstrated that this misconception is widespread[36]. We emailed authors in psychology, medicine and behavioural neuroscience inviting them to adjust a figure until they judged two means, with error bars, to be just statistically significantly different ($p<.05$). Few researchers answered within a correct range (17%, 24 of 140). The remainder were too conservative, setting the CIs to just touch or not overlap at all. The mean response positioned the CIs at the equivalent of $p = .009$.

The incorrect 'overlap' rule was also frequently applied to standard error bars, despite in this case leading to answers that were too lax. This further demonstrates the underlying confusion encountered in interpreting error bars. Whilst 95% CIs can overlap by roughly a quarter when the difference between the means is significant at $p\approx.05$, standard error bars need to be separated by about half of a standard error, or in other words one 'arm' of a standard error bar (Cumming & Finch, 2005). In the standard error group, about a quarter (25%, 44 of 179) of researchers answered within a correct range; the remainder gave answers that were too lax. The mean response in this group corresponded to $p=.109$.

10.4.2 Error bars for Independent Groups versus Error Bars for Repeated Measures

Not only did most researchers in the Belia et al. study fail to distinguish between the type of error bar presented (i.e., standard error or CI) they also failed to differentiate between the type of design used in the research scenario. Error bars around individual means are valid only for interpreting data from independent groups. For repeated measures data, the appropriate error bar is that around the difference between the two means. Only this bar removes the variability between individuals or sites, and allows us to look directly at the difference over time (e.g., the difference between pre and post impact). Only 11% (18 of 159) of researchers realised that bars around individual means could not used to determine statistical significance in repeated measures cases. As we noted: "It is a serious problem that the usual graphical conventions…do not make salient whether a factor is an across- or within-subjects factor." (Belia et al., in press).

---

[36] Schenker and Gentleman (2001) reported that the incorrect overlap rule is sometimes used by authors in medical and health journals. Payton, Greenstone and Schenker (2003) and Wolfe and Hanley (2002) also explain why the rule is incorrect.

10.4.3 The Confidence Level Misconception[37]

The Confidence Level Misconception is the mistaken belief that a 95% CI will on average capture 95% of replication means. In fact, a 95% CI will on average capture 83.4% of future replication means (with a skewed distribution). In a web-based study similar to the one described above, we (Cumming, Williams & Fidler, 2004) emailed researchers with an online task of estimating 10 future sample means from a given sample mean and a 95% CI.

Over three-quarters (78%, 105 of 134) of researchers placed 9 or 10 within the error bar limits. In open-ended comments many made clear that they had attempted to match the number of future means within the intervals to the confidence level (i.e., 95%). Corresponding results were found in the standard error group, with most researchers placing 6 or 7 of the future means within the standard error limits, again matching the number of means to the confidence level of standard errors (e.g. approximately 68%).

10.5 Obstacles to the Adoption of Confidence Intervals

Technical Developments

It is not only statistical cognition research that has neglected the study of CIs. Technical statistical advances also lag. How to calculate confidence intervals in complex, multivariate designs—and indeed how to construct graphs for such designs—is not straightforward and is in need of research and development. As I mentioned in Chapter Seven, Grayson, Pattison and Robins (1998) charged CI advocates with having relied "on oversimplified scenarios" (p.69). They were largely right, and the situation has only recently begun to be addressed. With the majority of statistical developments over the last half century being made within a significance testing framework, the scope of application for CIs remains comparatively narrow. Some exceptions to this include the work on CIs and non-central distributions (cited in Chapter Four and elsewhere) and, notably, Loftus and Masson (1994) and Masson and Loftus (2004), who discuss the selection of appropriate error terms for within-group CIs. Masson and Loftus (2004)

[37] Geoff Cumming named this the 'Confidence Level Misconception' however it may be better characterised as the 'Confidence Level Replication Misconception'. Furthermore, the study described may be controversial since the question itself does not actually have a correct answer. It requires thinking about a given realisation of a CI and it is perhaps a reasonable response to take the give realisation as a population model.

and Loftus (2002) also discuss graphical displays of CIs. Finch and Cumming (2001) and Cumming and Finch (2005) also provide advice on graphing and interpreting CIs.

Developing Appropriate Heuristics for Interpreting Confidence Intervals

In addition to increasing the scope of CI application, researchers and students have been offered little in the way of guidance for CI interpretation. As the studies in this chapter demonstrate, it would be foolish to assume CI interpretation is self-evident. The following 'rules of thumb' help may help researchers and students better interpret CIs. Cumming and Finch (2005) offer a more complete list of guidelines (or 'rules of eye').

- Each value in the interval as a plausible value for the parameter estimate. Consider the consequences of each value in the interval being the true value? (The upper and lower limits can be used as a shortcut in this process).
- The values within any interval are part of a distribution (the tails of the distribution sit just off the limits of the interval.) Values in the centre of the distribution, in the centre of the interval closest the point estimate, are more likely than values at either end.
- What effect would be psychologically/clinically/medically/biologically/ecologically important effect—as distinct from simply a statistically significant effect? Is the important effect in the interval? If so, where is it positioned?
- Note the width of the interval and what this tells you about the precision of the study. A very wide interval is indicative of low precision (similar to low statistical power). A narrow, focused interval is a sign of precise study.

10.6 Conclusion

The studies in Chapter Nine provide empirical evidence for some cognitive advantage of CIs. In particularly, these studies demonstrate that CIs help alleviate one the most serious misconceptions associated with NHST—that statistical non-significance is equivalent to 'no effect'. This is a major plus for CIs, since that misconception alone is responsible for much of the damage NHST has caused to the progress of science.

The studies in this chapter (Chapter Ten) warn of some misconceptions associated with CIs themselves. These are in some cases very different to the misconceptions associated with *p* values. For example, students' in these surveys were confused about the inferential nature of a CI, instead interpreting it as a descriptive statistics reminiscent of the range of individual scores. Results from other relevant studies we have conducted show that researchers too have misconceptions about CIs—including misinterpreting overlap of CIs, interpreting CIs from different designs in the same way and interpreting CIs and SE bars as the same thing.

On the balance of evidence, CIs would seem to have promise as an alternative to NHST. Yet, challenges remain. First, the application of CIs to complex designs has been largely neglected, with most advances over the last half a century being focused on NHST approaches. Second, despite the cognitive advantages CIs offer, they are far from being free of misconceptions. More efforts are need to study how researchers and students think about CIs, and how they are best presented and taught. The heuristics, or guidelines, for CI interpretation given here are a preliminary attempt at ensuring maximum information is gleaned from the reported statistic. Of course, for reform to be truly evidence-based such heuristics themselves require empirical research, as do other related proposals.

FUTURE DIRECTIONS

In the absence of adequate guidance from institutions such as the APA, and the absence of appropriate editorial pressure, statistical reform in psychology has an uncertain future. Despite the cogent arguments and dedicated efforts of so many, it remains hard to feel optimistic. As I write this, it has just been suggested that statistical power analysis and meta-analysis be cut from the NHST dominated statistics curricula in the psychology department I work in. The time has come—is perhaps well over-due—to provide an evidence base for statistical reform in psychology and other disciplines. This will entail providing empirical justification for adopting alternatives to statistical significance testing, and evidence-based guidance for implementing and interpreting those alternatives. An overarching goal of such research must be, in my opinion, to develop further links between two parallel literatures: the theoretical literature on human decision making and judgements under uncertainty (led by Gerd Gigerenzer, Daniel Kahneman and others) and the rhetorical literature of statistical reform.

As I have explained, in medicine, effect sizes and CIs were institutionalised through strict editorial policy in mid-1980s. It is often therefore assumed that editorial policy is the key to statistical reform for psychology. Yet my investigations of reporting practice in psychology journals empirically demonstrate that 'encouragements' are not capable of overcoming the inertia that exists. Furthermore, investigations in medicine demonstrate that reform is a complex process, and that changes in reporting practice are just the beginning. More collaborative efforts amongst journal editors and stricter enforcing of editorial recommendations in psychology may indeed affect some change. However, even medicine—with all its achievements—still has some way to go in freeing itself from the strange-hold of dichotomous decision making based on statistical significance. The change in reporting practice in medical journals was, however, non-trivial and there are important lessons for psychology in how the change was instituted. In particular, psychology, as a discipline, needs to overcome obstacles related to statistics education. Statistical textbooks for psychology students need to be re-written and new curricula developed. (I acknowledge of course the recent efforts in this area by Bruce Thompson, Mike Smithson and others.) Psychology also needs to confront the

conceptual issues related to a shift to estimation in discipline so often plagued by a lack of natural or at least universal measurement scales.

There has been a lot of repetition in this thesis already, and rather than offer another summary of the various arguments I will offer a series of remaining questions. Moving on from statistical significance is currently one of the most important items on the agenda of many social and life sciences. Answers to the questions below will be as important to economics, education, sociology, medicine and health sciences as they will be to psychology itself. Ecology in particular has much at stake in statistical reform. Ongoing misinterpretation of statistically non-significant results is a particular threat in the study of at-risk or endangered populations and there is much to gain from presentations of results that acknowledge uncertainty. The application of cognitive psychology in other scientific disciplines is becoming increasing common. The investigations entailed by these research questions offers yet another opportunity for psychology to make a worthwhile contribution.

First, do CIs free researchers from misconceptions associated with significance testing? The preliminary investigations I presented in Chapter Nine indicate promising results, but this work needs to be extended.

Second, what misconceptions are associated with CIs and what are their consequences? If researchers routinely make the errors identified in Chapter Ten, what effect will this have on the scientific literature or the progress of experimental disciplines? Are there other misconceptions not yet identified?

Third, how should CIs be presented and discussed? What are the formats that minimise misconceptions and maximise benefits? What guidelines and rules of thumb can be successfully taught and used?

Fourth, do CIs facilitate meta-analytic thinking? A brief examination of medical journals suggests that when CIs are used to conduct short meta-analyses before new experiments, the interpretation of new experimental results are qualitatively different—more insightful and imaginative. Encouraging researchers to incorporate prior information, and think beyond single experiments, is an important potential advantage of CIs. This question therefore deserves investigation.

Finally, if CIs and effect sizes are not a viable alternative to statistical significance tests, what might be better? The obvious next step is to ask such questions of Bayesian (and other) methods.

REFERENCES

Abelson, R.P. (1997). On the surprising longevity of flogged horses: why there is a case for the significance test. *Psychological Science, 8*, 12-15.

Abelson, R.P. (1954). Critical comment on "Learning and the principle of inverse probability." *Psychological Review, 61*, 276-278.

Abramson, L.Y., Seligman, M.E.P. & Teasdale, I. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology, 87*, 49-59.

Aiken, L., West, S., Sechrest, L., & Reno, R.R. (1990). The training in statistics, methodology, and measurement in psychology. *American Psychologist, 45,* 721–735.

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

Allison, D.B., Faith, M.S., & Gorman, B.S. (1996). Publication bias in obesity treatment trials? *International Journal of Obesity, 20*, 931-937

Altman, D.G. (2000a). Confidence intervals in practice. In D.G. Altman, D. Machin, T.N. Bryant, & M.J. Gardner (Eds.), *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed., pp. 6–14). London: BMJ Books.

Altman, D.G. (2000b). Statistics in medical journals: some recent trends. *Statistics in Medicine, 19*, 3275-3289.

Altman, D.G. (1991). Statistics in medical journals: developments in the 1980s. *Statistics in Medicine 10*, 1897–1913.

Altman, D.G. (1982a). 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. (1982b). 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. (1982c). Improving the quality of statistics in medical journals. In D.G. Altman & S.M. Gore (Eds.). *Statistics in Practice*. (pp. 21-24). London: BMJ Books.

Altman, D.G. & Gore, S.M. (Eds.). (1982). *Statistics in Practice*. London: BMJ Books.

Altman D.G., Machin D., Bryant T.N. & Gardner M.J. (2000). (Eds.) *Statistics with confidence: Confidence intervals and statistical guidelines*. (2nd ed.). London: BMJ Books.

Altman, M. (2004). Introduction. *Journal of Socio-economics, 33*, 615-630.

Ambroz, A, Chalmers, T.C. & Smith, H. (1978). Deficiencies of randomized controlled trials. *Clinical research 26*, 280.

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.R. & Burnham, K.P. (2002). Avoiding pitfalls when using information theoretic methods. *Journal of Wildlife Management, 66,* 912-918.

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

Andrews, L. & Leopold, B.D. (1999). Guidelines for authors and reviewers of *Wildlife Society Bulletin* manuscripts. Retrieved on 01-11-05 from: (http://www.wildlife.org/publications/bulletinguidelines.pdf)

Bailar, J.C. & Mosteller, F. (Eds.). (1986). *Medical Uses of Statistics*. Waltham, MA, USA: NEJM Books.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation.* San Francisco: Jossey-Bass.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Bandt, C.L. & Boen, J.R. (1972). A prevalent misconception about sample size, statistical significance, and clinical importance. *Journal of Periodontology, 43*, 181–183.

Barber, B. (1961). Resistance by scientists to scientific discovery, *Science, 134*, 596-602.

Bauchau, V. (1997). Is there a "file drawer problem" in biological research? *OIKOS, 19*, 407–409.

Beal, S.L. (1989). Sample size determination for confidence intervals on the population mean and on the differences between two population means. *Biometrics, 45*, 969 -977.

Beauchamp, K.L., & May, R.B. (1964). Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports, 14*, 272.

Beutler, L.E., & Moleiro, C. (2001). Clinical versus reliable and significant change. *Clinical Psychology: Science & Practice, 8,* 441–445.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (in press). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*.

Ben-Haim, Y. (2001). *Information-Gap Decision Theory: Decisions Under Severe Uncertainty*. Cornwall, UK: Academic Press.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536.

Berlin, J.A., Begg, C.B. & Louis, T.A. (1989). An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association, 84*, 381-392.

Bernstein, B.B. & Zalinski, J. (1983). An optimal sampling design and power tests for environmental biologist. *Journal of Environmental Management, 65*, 35–43.

Bezeau, S., & Graves, R.E. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23*, 399–406.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70*, 107-115.

Bland, J.M. & Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal, 310*, 170.

Boivin, M., Pérusse, D., Dionne G, Saysset, V., Zoccolillo, M., Tarabulsy, G.M., Tremblay, N. & Tremblay, R.E. (2005). *Journal of Child Psychological Psychiatry, 46*, 612-630.

Bookstein, F.L. (1998). Statistical significance testing was not meant for weak corroborations of weaker theories. *Behavioral and Brain Sciences, 21*, 195-196.

Borak, J. & Veilleux, S. (1982). Errors in intuitive logic among physicians. *Social Sciences and Medicine, 16*, 1939-1947

Borenstein, M. (1997). Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy, Asthma, & Immunology, 78*, 5-16.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information, 14*, 20-47

Bozarth, J.D., & Roberts, R.R. (1972). Signifying significance. *American Psychologist, 27*, 774-775.

Budge, G., & Katz, B. (1995). Constructing psychological knowledge: Reflections on science, scientists and epistemology in the APA *Publication Manual*. *Theory & Psychology*, *5*, 217-231.

Burgman, M.A. (2005). *Risks and decision for conservation and environmental management.* Cambridge: Cambridge University Press.

Burgman, M.A., Maslin, B., Andrewartha, D., Keatley, M., Boek, C. & McCarthy, M. (2000). Detecting trends in sighting data: power and an application to Western Australian Acacia species. In S. Ferson & M. A. Burgman (Eds.). *Quantitative methods for conservation biology*. (pp. 7-26). New York, USA: Springer-Verlag.

Burgman, M. A. & Possingham, H. P. (2000). Population viability analysis for conservation: the good, the bad and the undescribed. In A.G. Young & G.M. Clarke (Eds.). *Genetics, demography, and viability of fragmented populations*. (pp 97-112). Cambridge: Cambridge University Press.

Burnham, K.P. & Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, USA: Springer-Verlag.

Campbell, J.P. (1982). Editorial: some remarks from the outgoing editor. *Journal of Applied Psychology, 67*, 691-700

Carey, J.M. & Keough, M. (2002). The variability of estimates of variance, and its effect on power analysis in monitoring design. *Environmental Monitoring and Assessment, 74*, 225-241.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Chambless, D.L., & Hollon, S.D. (1988). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–19.

Chandler, R.E. (1957). The statistical concepts of confidence and significance. *Psychological Bulletin, 54*, 429-430

Chase, L.J. & Chase, R.B. (1976). A statistical power analysis of applied psychology research. *Journal of Applied Psychology, 61*, 234–237.

Cherry, S. (1998). Statistical tests in publications of The Wildlife Society. *The Wildlife Society Bulletin, 26*, 947-953.

Cherry, S. (1996). A comparison of confidence interval methods for habitat use-availability studies. *Journal of Wildlife Management, 60*, 653-658.

Chorpita, B. (2002). The tripartite model and dimensions of anxiety and depression: an examination of structure in a large school sample. *Journal of Abnormal Child Psychology, 30*, 177-190.

Chow, S.L. (2002). Issues in Statistical Inference. *History and Philosophy of Psychology Bulletin*, *14*, 30-41.

Chow, S.L. (2000). The Popperian framework statistical significance, and rejection of chance. *Behavioral and Brain Sciences, 23*, 294-298.

Chow, S. L. (1998). A précis of "Statistical Significance: Rationale, Validity and Utility." *Behavioral and Brain Sciences*, *21*, 169-194.

Chow, S.L. (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks, CA, USA: Sage.

Chow, S.L. (1991). Some reservations about statistical power. *American Psychologist*, *46*, 1088-1089.

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105-110.

Chow, S.L. (1987). Meta-analysis of pragmatic and theoretical research: A critique. *Journal of Psychology*, *121*, 259-271.

Clark-Carter, D. (1997). The account taken of statistical power in research journal in the British Journal of Psychology. *British Journal of Psychology, 88*, 71-83.

Clark, J & Lavine, M. 2001. Bayesian statistics in ecology. In S.M. Scheiner & J. Gurevitch (Eds.). *Design and Analysis of Ecological Experiments*. (pp. 327-346). Oxford, UK: Oxford University Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. (1990). Things I have learned so far. *American Psychologist, 45*, 1304-1312.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*, 107-112.

Cohen, J. (1970). Approximate power and sample size determination for common one-sample and two sample hypothesis tests. *Educational and Psychological Measurement, 30*, 811-831.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Coulson, M., Fidler, F. & Cumming, G. (2005). *Understanding of confidence intervals by researchers in psychology, behavioural neuroscience, and medicine.* **In preparation.**

Cowles, M. (1989). *Statistics in psychology: an historical perspective*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Craig, J.R., Eison, C.L., & Metze, L.P. (1976). Significance tests and their interpretation: An example utilizing published research and omega-squared. *Bulletin of the Psychonomic Society, 1*, 280-282.

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116-127.

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

Cumming G. & Finch S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and non-central distributions. *Educational and Psychological Measurement, 61*, 532-574.

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

Curran-Everett, D., Taylor, S., & Kafadar, K. (1998). Fundamental concepts in statistics: elucidation and illustration. *Journal of Applied Physiology, 85*, 775–786.

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.

Daly, L. (2000). Confidence intervals and sample sizes. In D.G. Altman, D. Machin, T.N. Bryant, & M. J. Gardner (Eds.), *Statistics with confidence* (2nd ed., pp. 139–152). London: BMJ Books.

Danziger, K. (1990). *Constructing the subject: historical origins of psychological research*. Cambridge: Cambridge Press.

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Kruger. G. Gigerenzer & M. Morgan. (Eds.), *The probabilistic revolution. Vol. 2: Ideas in the sciences*. (pp. 35-47). Cambridge, MA: MIT

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42*,145-151.

Dar, R., Serlin, R.C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75–82.

Dawes, R.M. (1988). Rational choice in an uncertain world. NewYork: Harcourt, Brace, Javonovich.

Demming, W.E. (1974) Invited talk at Princeton University on November 17 [cited in Salsburg, D. (2001). *The lady tasting tea: how statistics revolutionized science in the twentieth century*. New York, USA: W.H. Freeman and Company.]

Diaconis, P. & Freedman, D. (1981). The persistence of cognitive illusions. *Behavioral and Brain Sciences, 4*, 378-399

DiStefano, J. (2003). A confidence interval approach to data analysis. *Forest Ecology and Management, 187*, 173-183.

DiStefano J., Fidler F. & Cumming G. (2005). Effect size estimates and confidence intervals: An alternative focus for the presentation and interpretation of ecological data. In A.R. Burk (Ed.) *New Trends in Ecology Research.* Hauppauge, NY, USA: Nova Science Publishing,

Earleywine, M. (1993). The file drawer problem in the meta-analysis of subjective responses to alcohol. *American Journal of Psychiatry, 150*, 1435–1436.

Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Elith, J., & Burgman, M.A. (2003). Habitat models for population viability analysis. In C.A. Brigham & M.W. Schwartz (Eds.). (pp. 203-235). *Population viability in plants*. Springer-Verlag, Berlin.

Ellison, A.M. (2004). Bayesian inference in ecology. *Ecology Letters, 7*, 509-520.

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

Eysenck, H.J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review, 67*, 269-271.

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

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*, 75-98.

Food and Drug Administration (USA). Centre for Drug Evaluation and Research. *Guidance for Industry, E9 Statistical Principles for Clinical Trials*. Retrieved on 18-02-05 from: (http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF).

Feinstein, A.R. 1974. Clinical biostatistics XXV A survey of the statistical procedures in general medical journals. *Clinical Pharmacology Therapy, 15, 97-107.*

Ferson, S. (2005). *Bayesian methods in risk assessment.* Unpublished report prepared for the Bureau de Recherches Geologiques et Minieries (BRGM).

Feynman, R. (1967). *The character of physical law*. Cambridge, MA, USA: MIT Press

Fidler, F., Cumming, G., Burgman, M.A., Buttrose, R. & Thomason, N. (In press). Have criticisms of null hypothesis significance testing had an impact on conservation biology. *Conservation Biology*.

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

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C. & Schmitt, R. (2005). Toward improved statistical reporting in 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., et al. (2004). Reform of statistical inference in psychology: The case of memory and cognition. *Behavior Research Methods, Instruments, & Computers, 36,* 312–324.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future APA guidelines for statistical practice. *Theory and Psychology, 12*, 825–853.

Fisher, R.A. (1973). *Statistical methods and scientific inference***.** New York, USA: Hafner.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.

Fisher, R.A. (1925). *Statistical methods for research workers*. London, UK: Oliver and Boyd.

Fisher, R.A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *Journal of agricultural science, 11*, 107-135.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52*, 399-433.

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.

Frai, J.L., Nielsen, S.E., Merrill, E.H., Lele, S.R., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B. & Beyer, H.L. (2002). Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology, 41*, 201-212.

Fraley, R. (2003) *End of the semester thoughts on the significance testing debate*. Retrieved on 02-02-05 from: (http://www.uic.edu/classes/psych/psych548/fraley/NHSTsummary.htm)

Freedman, L. (1996). Editorial. Bayesian statistical methods: A natural way to assess clinical evidence. *British Medical Journal, 313*, 569-570.

Freiman, J.A., Chalmers, T.C., Smith, J.H. & Kuebler, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized clinical trials. Survey of 71 "negative" trials. *New England Journal of Medicine, 299*, 690-4.

Frick, R.W. (1998). Chow's defense of null-hypothesis testing: too traditional. *Behavioral and Brain Sciences, 21*, 199.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1,* 379-390.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, *70*, 245-251.

Funk, W.C., Donnelly, M.A. & Lips, K.R. (2005). Alternative views of amphibian toe-. clipping. *Nature, 433*,193.

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

Gardner, M.J. Altman, D.G., Jones, D.R. & Machin, D. (1983). Is the statistical assessment of papers submitted to the *British Medical Journal* effective? *British Medical Journal (Clinical Research Edition), 286*, 1485-8.

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. (1997). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200.

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, USA: Lawrence Erlbaum.

Gigerenzer, G. (1989). The tools-to-theories hypothesis: on the art of theory construction in cognitive psychology. In *24th International Congress of Psychology of the International Union of Psychological Science, 4*, 163-171. Oxford, UK.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Kruger, J. Lorraine & G. Gigerenzer (Eds.). *The Probabilistic Revolution, Vol. 2: Ideas in the sciences* (pp. 11-33.). Cambridge, MA, USA: The MIT Press.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics*. Hillsdayle, NJ, USA: Lawerence Erlbaum.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. New York, USA: Chapman & Hall.

Gladis, M.M., Gosch, E.A., Dishuk, N.M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 320–331.

Glaser, D.N. (1996). The need for a moratorium on significance testing. *Journal of cardiovascular nursing, 11*, viii-ix.

Glass, G.V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Good, I.J. (1982). Comment. *Journal of the American Statistical Association, 77*, 342-344.

Goodman, S.N. & Berlin, J.A. (1994). The Use of Predicted Confidence Intervals when Panning Experiments and the Misuse of Power When Interpreting Results. *Annals of Internal Medicine, 121*, 200-206.

Gore, S, M. (1982). Assessing methods: Art of significance testing. In D.G. Altman & S.M. Gore (Eds.). *Statistics in Practice*. (pp. 70-73). London: BMJ Books.

Gore, S.M., Jones, I.G. & Rytter, E.C. (1977) Misuse of statistical methods: critical assessment of articles in the BMJ from January to March 1976. *British Medical Journal*, *1*, 85–87

Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*, 54-61

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. *Lancet, 2*, 538.

Green, R.H. (1989). Power analysis and practical strategies of environmental monitoring. *Environmental Research, 50*, 195-205.

Greenland, S. (1998). Meta-analysis. In K.J. Rothman & S.Greenland (Eds.). *Modern epidemiology* (2nd ed., pp. 643–673). Philadelphia, USA: 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.

Grieve, A.P. (1991). Confidence intervals and sample sizes. *Biometrics, 47*, 1597-602.

Grove, W.M., & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Guerra, R., Etzel, C.J., Goldstein, D.R. & Sain, S.R. (1999). Meta-analysis by combining p-values: simulated linkage studies. *Genetic Epidemiology, 17(Suppl 1),* S605-9.

Guion, R.M. (1983). Editorial: Comments from the new editor. *Journal of Applied Psychology, 68*, 547-551.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge, U.K: Cambridge University Press.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.

Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York, USA: Wiley.

Hall, P., & B. Selinger. (1986). Statistical significance: balancing evidence against doubt. *Australian Journal of Statistics* 28:354-370.

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

Hammond, G. (1996). The objections to null hypothesis testing as a means of analysing psychological data. *Australian Journal and Psychology, 48*, 104–106.

Harlow, L.L. (1997). Significance Testing in Introduction and Overview. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds). *What If There Were No Significance Tests?* (pp.1-17). Mahwah, NJ, USA: Lawrence Erlbaum.

Harris, E.K. (1993). On *p* values and confidence intervals (why can't we *p* with more confidence?). *Clinical chemistry, 39*, 927-928.

Harris, R. (1998). "With friends like these…" Three flaws in Chow's defence of significance testing. *Behavioral and Brain Sciences, 21*, 202-203.

Harris, R. (1997). Significance tests have their place. *Psychological Science*, *8*, 8-11.

Harwood, J. (2000). Risk assessment and decision analysis in conservation. *Biological Conservation, 95*, 219–226.

Hayes, J.P. & Steidl, R.J. (1997). Statistical power and analysis and amphibian population trends. *Conservation Biology, 11*, 273–275.

Henderson, H.R. (1993). Chemistry with confidence: should clinical chemistry require confidence intervals for analytical and other data? *Clinical chemistry, 39*, 929-935.

Hilborn, R. & Mangel, M. (1997). *The ecological detective: Confronting models with data.* Princeton, USA: Princeton University Press

Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine, 58*, 295-300.

Hill, A.B. (1937). *Principles of medical statistics*. London, UK: Lancet.

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, USA: Kluwer.

Hiroto, D.S. & Seligman, M.E.P. (1975). Generality of learned helplessness in man. *Journal of Personality and Social Psychology, 31*, 311-327.

Hogben, L. (1957) *Statistical Theory: The relationship of probability, credibility, and error; an examination of the contemporary crisis in statistical theory from a behaviourist viewpoint*. London, UK: Allen & Unwin.

Hoenig 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.

Hollon, S.D. & Flick, S.N. (1988). On the meaning and methods of clinical significance. *Behavioral Assessment, 10*, 197–206.

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

Hubbard, R. & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology--And its future prospects. *Educational and Psychological Measurement*, *60*, 661-681.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317-333.

Huberty, C.J., & Pike, C.J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.). *Advances in Social Science Methodology (Vol. 5)* (pp. 1-22) Stamford, CT, USA: JAI Press.

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

Hunter, J.E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA, USA: Sage.

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

Hutton, J.L., Cooke, T. & Pharoah, P.O.D. (1994). Life expectancy in children with cerebral palsy. *British Medical Journal, 309*, 431 -435.

Institute for Scientific Information (ISI) (2004). *Journal Citation Reports for Social Science.*

Institute for Scientific Information (ISI) (2004). *Journal Citation Reports for Science.*

International Committee of Medical Journal Editors. (1988a). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine, 108*, 258–265.

International Committee of Medical Journal Editors. (1988b). Uniform requirements for manuscripts submitted to biomedical journals. *British Medical Journal, 296*, 401–408.

Iyengar, S., & Greenhouse, J.B. (1988). Selection models and the file-drawer problem. *Statistical Science, 3*, 109–135.

Jacobson, N.S., Follette, W.C. & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.

Jacobson, N.S., Roberts, L.J., Berns, S.B. & McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300–307.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Jeffreys, H. (1961). Theory of probability. (3rd Ed.) Oxford, UK: Oxford University Press.

John, I. D. (1992). Statistics as rhetoric in psychology. *Australian Psychologist, 27*, 144-149.

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.

Johnson, D.H. (1999) The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*, 763-772.

Kadane, J. B. (1995). Prime time for Bayes. *Controlled Clinical Trials, 16*, 313-318.

Kaiser, H.F. (1960). Directional statistical decisions. *Psychological Review, 67*, 160-167.

Kazdin, A.E. (1999). The meanings and measurement of clinical significance. Journal of Consulting and Clinical Psychology, 67, 332–339.

Kelley, K. (2005). The effects of non-normal distributions on confidence intervals around the standardised mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement, 65,* 51-69.

Kelley, K., Maxwell, S.E. & Rausch, J.R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation and the Health Professions, 26*, 258-287.

Kempthorne, O. (1976). Of what use are tests of significance and tests of hypotheses. *Communications in statistics, Series A 5*, 763-777.

Kendall, M. G. (1942). On the future of statistics. *Journal Royal Statistical Society, 105*, 69-80.

Kendall, P. (1957). Note on Significance Tests. In R.K. Merton, G.C. Reader & P. Kendall (Eds.). *The Student Physician*. (pp.301-305). Cambridge, MA, USA: Harvard University Press.

Kendall, P.C. (1997). Editorial. *Journal of Consulting and Clinical Psychology, 65,* 3–5.

Kendall, P.C. & Grove, W. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment, 10,* 147–158.

Kendall, P.C., Marrs-Garcia, A., Nath, S.R. & Sheldrick, R.C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285–299.

Keppel, G. (1968). Retroactive and proactive inhibition. In T.R. Dixon & D.L. Horton. *Verbal Behavior and General Behavior Theory* (pp. 172–213.) Englewood Cliffs, NJ: Prentice-Hall.

Kieffer, K.M., Reese, R.J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69,* 280–309.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213-218.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kirk, R.E. (Ed.). (1972). *Statistical issues: A reader for the behavioral sciences*. Monterey, CA, USA: Brooks/Cole.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review, 24*, 328-338.

Kline, R.B. (2004). *Beyond significance testing: reforming data analysis methods in behavioral research*. Washington DC, USA: American Psychological Association.

Knapp, S. & Jackson, D. (2001, August). *New edition of the APA* Publication Manual. Workshop (Session 2129) presented at the annual meeting of the American Psychological Association, San Francisco.

Kosciulek, J. F., & Szymanski, E. M. (1983). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36*, 212-219.

Kraemer, H.C. (1992). Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. *Psychoneuroendocrinology, 17*, 527-536.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*, 1372-1381.

Krüger, L., Gigerenzer, G. & Morgan, M.S. (Eds.). (1987). *The Probabilistic Revolution, vol. 2. Ideas in the Sciences.* Cambridge, MA/London: MIT Press.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*. Chicago, USA: University of Chicago Press.

LaForge, R. (1967). Confidence intervals or tests of significance in scientific research? *Psychological Bulletin, 68*, 446-447.

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.). *The Methodology of Scientific Research Programs.* (pp. 8-93). Cambridge, UK: Cambridge University Press.

Lande, R. (1988). Demographic models of the Northern Spotted Owl (Strix occidentalis caurina). *Oecologia, 75*, 601–607.

Langman, M.J.S. (1986) Towards estimation and confidence intervals. *British Medical Journal*, *292*, 716.

Lashley, B (1998). A defense of statistical power analysis. *Behavioral and Brain Sciences, 21*, 209-210

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

Lees, M.C. & Neufeld, R.W.J. (1994). Matching the limits of clinical inference to the limits of quantitative methods: A formal appeal to practice what we consistently preach. *Canadian Psychology, 35,* 268–282.

Lewis, D. & Burke, C.J. (1949).The use and misuse of the chi-square test. *Psychological Bulletin, 46*, 433-489.

Lindgren, B.R., Wielinski, C L., & Finkelstein, S.M. (1994). Contrasting clinical and statistical significance within the research setting. *Pediatric Pulmonology, 18,* 64–65.

Lindquist, E.F. (1940). *Statistical analysis in educational research*. Boston, USA: Houghton Mifflin

Lindquist, E.F. (1938). *A first course in statistics, their use and interpretation in education and psychology*. Boston, USA: Houghton Mifflin.

Lockhart, R.S. (1998). *Introduction to Statistics and Data Analysis*. New York, USA: W.H. Freeman.

Loftus, G.R. (2002). Analysis, interpretation, and visual presentation of experimental data. In J. Wixted & H. Pashler (Eds.), *Stevens' handbook of experimental psychology*: *Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 339–390). New York, USA: Wiley.

Loftus, G.R. (1996). Why psychology will never be a real science until we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161-171.

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

Loftus, G.R. (1993b). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments & Computers*, *25*, 250-256.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36*, 102-104.

Loftus, G.R., & Masson, M. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Lovie, A.D. (1979). The analysis of variance in experimental psychology: 1934-1945. *British Journal of Mathematical and Statistical Psychology, 32*, 151-178.

Luus, H.G., Muller, F.O. & Meyer, B.H. (1989). Statistical significance versus clinical relevance: II. The use and interpretation of confidence intervals. *South African Medical Journal, 76,* 626–629.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151-159.

McCarthy, M.A. (in press). *Bayesian methods for ecologists*. Cambridge University Press.

McCarthy, M.A. & Parris, K.M. (2004). Clarifying the effect of toe clipping on frogs with Bayesian statistics. *Journal of Applied Ecology, 41*, 780-786.

McCloskey, D.N. (1995). The insignificance of statistical significance. *Scientific American, 272*, 32-33.

McCloskey, D.N. (1992). The bankruptcy of statistical significance. *Eastern Economic Journal, 18,* 359-361.

Maddock, J.E. & Rossi, J.S. (2001). Statistical power of articles published in three health psychology related journals. *Health Psychology, 20*, 76-78.

Mainland, D. (1984). Statistical rituals in clinical trials: Is there a cure? British Medical Journal, 288, 341–343.

Mainland, D. (1952). *Elementary medical statistics; the principles of quantitative medicine*. Philadelphia, USA: WB Saunders.

Manchanda, R. (1986). Criteria for measuring change: Statistical significance versus clinical significance. *British Journal of Psychiatry, 148,* 744–745.

Mantha, S., Thisted, R., Foss, J., Ellis, J.E. & Roizen, M.F. (1993). A proposal to use confidence intervals for visual analog scale data for pain measurement to determine clinical significance *Anesthesia & Analgesia, 77*, 1041-7.

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. (1997). *The progress of experiment: science and therapeutic reform in the United States, 1900-1990*. New York, USA: Cambridge University Press.

Marks, R.G., Dawson-Saunders, E.K., Bailar, J.C., Dan, B.D. & Verran, J.A. (1988) Interactions between statisticians and biomedical journal editors. *Statistics in Medicine*, *7*, 1003–1011.

Masson, M. & Loftus, G.R. (2003). Using confidence intervals for graphically based interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

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

May, R.M. (2004). Ethics and amphibians. *Nature, 431*, 403

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

Mayo, D.E. (1998). Some problems with Chow's problems with power. *Behavioral and Brain Sciences, 21*, 212-213.

Meehl, P.E. (1998). The power of quantitative thinking. Speech delivered upon receipt of the James McKeen Cattell Fellow award at the meeting of the American Psychological Society, Washington, D.C., May 23. Retrieved on 20-10-05 from: (http://www.tc.umn.edu/~pemeehl/).

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.

Meehl, P.E. (1967). Theory testing in psychology and physics: a methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P.E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4*, 268-273.

Meehl, P.E. (1954). *Clinical vs. statistical prediction: a theoretical analysis and a review of the evidence*. Minneapolis, MN, US: University of Minnesota Press.

Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Miller, W.R. & Seligman, M.E.P. (1975). *Depression and learned helplessness in man. Journal of Abnormal Psychology, 84,* 228-238.

Miller, W.R. & Seligman, M.E.P. (1973). Learned helplessness, depression and the perception of reinforcement. *Journal of Abnormal Psychology, 82*, 62-73.

Mone, M.A., Mueller, G.C. & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49*, 103-120.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Morrison, D.E., & Henkel, R.E. (1969). Significance tests reconsidered. *American Sociologist, 4*, 131-140.

Mulaik, S.A., Raju, N.S. & Harshman, R.A.. (1997). There is a time and a place for significance testing. In S.A.Mulaik., L.L. Harlow, & J.H. Steiger (Eds.). *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ, USA: Lawrence Erlbaum.

Muris, P., Schmidt, H., Merckelbach, H., & Schouten, E. (2001). The structure of negative emotions in adolescents. *Journal of Abnormal Child Psychology, 29*, 331-337.

Murphy, K.R. (1997). Editorial. *Journal of Applied Psychology, 82,* 3–5.

Nester, M.R. (1998). Significance tests cannot be justified in theory-corroboration experiments. *Behavioral and Brain Sciences, 21*, 213-213.

Newcombe, R.G. & Altman, D.G. (2000) Proportions and their differences. In D.G. Altman, D. Machin, T.N. Bryant & M.J. Gardner (Eds.) *Statistics with Confidence* (pp. 39-50). London: BMJ Books.

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

Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical transactions of the Royal Society, A231*, 289-337.

Neyman, J. & Pearson, E.S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika, 20*, 174-240

Neyman, J. & Pearson, E.S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika, 20*, 263-294.

Nicewander, W.A. & Price, J.M. (1978). Dependent variable reliability and the power of significance tests. *Psychological Bulletin, 85*, 405-409.

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

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20*, 641-650.

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

Ogles, B.M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review, 21,* 421–446.

Otis, D. (1995). Journal News. *Journal of Wildlife Management, 59*, 630.

Parris, K.M., McCarthy, M.A. (2001). Identifying effects of toe clipping on anuran return rates: the importance of statistical power. *Amphibia Repilia, 22*, 275-289.

Payton, M.E., Greenstone, M.H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science, 3,* Article 34..Retrieved May 16 2005 from www.insectscience.org/3.34

Pearlman, K., Schmidt, F.L. & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology, 65,* 373-407.

Pearson, E.S. (1990). *'Student', A statistical biography of William Sealy Gosset*, In R. L. Plackett with the Assistance of G. A. Barnard, Oxford: University Press.

Pedhazur, E.J. & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, N.J., USA: Erlbaum.

Peterman, R.M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences, 47*, 2–15.

Petro, R. & Doll, R. (1977). When is significant not significant? *British Medical Journal, ii,* 259.

Pocock, S.J., Hughes, M.D. & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials. *New England Journal of Medicine, 317***,** 426-32.

Pollard, P. (1993). How significant is "significance?" In G.Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 448-460.) Hillsdale, N.J, USA: Lawrence Erlbaum.

Pollard, P. & Richardson, J.T.E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102*, 159-163.

Popper, K. (1969). *Conjectures and Refutations.* London: Routledge and Kegan Paul.

Popper, K. (1968). *The Logic of Scientific Discovery.* London: Hutchinson.

Postman, L., Stark, K. & Fraser, J. (1968). Temporal changes in interference. *Journal of Verbal Learning and Verbal Behavior, 7*, 672–694.

Pruzek, R.M. (1997). An introduction to Bayesian inference and its applications. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds). *What If There Were No Significance Tests?* (pp. 287-318). Mahwah, NJ, USA: Lawrence Erlbaum.

Reed J.M. & Blaunstein A.R. (1995). Assessment of "nondeclining" amphibian populations using power analysis. *Conservation Biology, 9*, 1299-1300.

Reichardt, C.S., & Gollob, H.F. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*, 117-128.

Reichardt, C.S., & Gollob, H.F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds). What if there were no significance tests? (pp. 259-284). Mahwah, NJ, USA: Lawrence Erlbaum.

Relman, A.S (1986). Preface. In J.C. Bailar & F. Mosteller (Eds.). *Medical Uses of Statistics*. Waltham, MA, USA: NEJM books.

Rennie, D. (1978). Vive la difference ($p < 0.05$). *New England Journal of Medicine, 299*, 828–829.

Richards, J.M. (1982). Standardized versus unstandardized regression weights. *Applied Psychological Measurement, 6,* 201-212.

Rigby, A.S. (1999). Getting past the statistical referee: moving away from *P*-values and towards interval estimation. *Health Education Research, 14*, 713-715.

Rigby, A.S. (1998). Statistical methods in epidemiology. I. Statistical errors in hypothesis testing. *Disability and Rehabilitation, 20*, 121–126.

Rindskopf, D.M. (1998). Null-hypothesis tests are not completely stupid, but Bayesian statistics are better. *Behavioral and Brain Sciences, 21*, 215-216.

Rindskopf, D.M. (1997). Testing "small," not null, hypotheses: classical and Bayesian approaches. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds). *What If There Were No Significance Tests?* (pp. 319-332). Mahwah, NJ, USA: Lawrence Erlbaum.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R. (1992). Effect size estimation, significance testing, and the file-drawer problem. *Journal of Parapsychology, 56*, 57-58.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Beverly Hills, CA, USA: Sage.

Rosenthal, R. (1984). *Meta-analytic Procedures for Social Research*. Beverly Hills, CA, USA: Sage Publications

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology, 51,* 4–13.

Rosenthal, R. (1979). The ‘‘file drawer problem’’ and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosenthal, R., & DiMatteo, M.R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of significance. *Psychological Reports, 15*, 570

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5,* 329–334.

Rosenthal, R. & Rubin, D.B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106*, 332-337.

Rosenthal, R. & Rubin, D.B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.

Rosenthal, R., & Rubin, D.B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rosnow, R.L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57,* 221–237.

Rosnow, R.L & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Rosnow, R.L. & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology, 35*, 203-208.

Rossi, J. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 175-198). Mahwah, NJ, USA: Lawrence Erlbaum.

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

Rothman, K.J. (1986) Significance questing (editorial). *Annals of Internal Medicine, 105*, 445-447.

Rothman, K.J. (1978a). Estimation of the confidence limits for the cumulative probability of survival in life table analysis. *Journal of Chronic Diseases, 31*, 557–560.

Rothman, K.J. (1978b). A show of confidence. *New England Journal of Medicine, 299*, 1362–1363.

Rothman, K.J. (1975). Computation of exact confidence intervals for the odds ratio. *International Journal of Bio-Medical Computing, 6*, 33–39.

Royall, R. (1986). The effect of sample size on the meaning of significance tests. *American Statistician*, *40*, 313-315.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-392). Mahwah, NJ, USA: Lawerence Erlbaum.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416-428. [Reprinted in D.E. Morrison & R.E. Henkel (Eds.). (1970). *The Significance Testing Controversy* (pp. 216-231). Chicago, USA: Aldine.

Rucci, A.J. & Tweney, R.D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin, 87*, 166-184.

Salsburg, D. (2001). *The lady tasting tea: how statistics revolutionized science in the twentieth century*. New York, USA: W.H. Freeman and Company.

Savage, L.J. (1961). The Foundations of Statistics Reconsidered. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, I*, 583.

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.

Sawyer, A.G., & Peter, J.P. (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research, 20*, 122-133.

Schafer, W.D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education, 61*, 383-387.

Schenker, N., & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*, 182-186.

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. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F.L. Berner, J.G. & Hunter, J.E. (1973). Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology, 58*, 59.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In L. Harlow, S. Mulaik, & J. Steiger (Eds.). *What if there were no significance tests?* (pp. 37–63). Mahwah, NJ, USA: Lawerence Erlbaum.

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

Schmidt, F.L. & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. (1976), Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61*, 473-485.

Schneeweiss, S., Maclure, M., Carleton, B.C., Glynn, R.J. & Avorn, J. (2004). Clinical and economic consequences of a formulary restriction of nebulized respiratory drugs in adults: Direct comparison of randomized and observational evaluations. *British Medical Journal, 328*, 560-567.

Schor, S. & Karten, I. (1966) Statistical evaluation of medical journal manuscripts. *Journal of the American Medical Association, 195*, 145–150.

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*, 134–143.

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

Seldrup J. (1997). Whatever happened to the t-test? *Drug Information Journal, 31*, 745-750.

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

Seligman, M.E.P. (1975). *Helplessness: On Depression, Development, and Death*. New York: W.H. Freeman.

Seligman, M.E.P. (1972). Learned helplessness. *Annual Review of Medicine, 23*, 407-412.

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

Selvin, H.C. (1957). A critique of tests of significance in survey research. *American Sociological Review, 22, 519-527.*

Serlin, R.C. (1993).Confidence intervals and the scientific method: a case for Holm on the range. *Journal of Experimental Education, 61*, 350-360.

Serlin, R.C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology*, *34*, 365-371.

Serlin, R.C. & Lapsey, D.K. (1985).Rationality in psychological research: the good-enough principle. *American Psychologist, 40*, 73-83.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*, 293-316.

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

Signorelli, A. (1974). Statistics: tool or master of the psychologist? *American Psychologist, 11*, 221-223.

Skinner, B.F. (1984). Methods and theories in the experimental analysis of behavior. *Behavioral and Brain Sciences, 7*, 524-525.

Skinner, B.F. (1971). *Beyond freedom and dignity*. New York, USA: Alfred Knopf.

Skinner, B.F. (1969). *Contingencies of reinforcement*. New York, USA: Appleton Century Crofts.

Skinner, B.F. (1956). A case history in scientific method. *American Psychologist, 11*, 221-223.

Smith, M., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*,752-760.

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

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605-632.

Smithson, M.J. (2000). *Statistics with confidence*. London: Sage.

Snedecor, G.W. (1937). *Statistical methods*. Ames, IA: Collegiate Press.

Snyder, P. & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*, 334-349.

Sober, E. (2005). Sex Ratio Theory, Ancient and Modern—the debate among Arbuthnot, Bernouilli, and DeMoivre, and the evolutionary ideas of Darwin, Dusing, Fisher, Williams, and Hamilton. In J. Riskin (Ed.). *The Sistine Gap: Essays on the History and Philosophy of Artificial Life.* **In preparation.**

Sokal, R.R. & Rohlf, F.J. (1969). *Biometry*. San Fransisco, CA, USA: W.H. Freeman and company.

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association, 84*, 608-610.

Spiegelhalter, D.J. (2004). On wasps and club dinners. *Significance, 1*, 183.

Spiegelhalter, D.J., Best, N.G., Carlin, B.R. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 64*, 583–616.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R. & Abrams, K.R. (1999). Methods in health service research: an introduction to Bayesian methods in health technology assessment. *British Medical Journal, 319*, 508–12.

Spiegelhalter D.J., Freedman L.S. & Parmar M.K.B. (1994). Bayesian approaches to randomised trials. *Journal of the Royal Statistical Society Association, 157*, 357.

SPSS Inc. (2004). SPSS for Windows [Computer software] (Version Release 13.0). Chicago: Author.

Stam, H.J. & Pasay, G.A. (1998). The historical case against null-hypothesis significance testing. *Behavioral and Brain Sciences, 21*, 219-220.

Steiger, J.H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182.

Steiger, J.H., & Fouladi, 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). Mahwah, NJ, USA: Lawrence 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.

Stigler, S.M, (1999) *Statistics on the table: the history of statistical concepts and methods*. Harvard University Press, Cambridge, Massachusetts.

Stigler, S.M. (1986). *The history of statistics: the measurement of uncertainty before 1900*. Cambridge, MA, USA: Belknap/Harvard University Press.

Stigler, S.M. (1990). A Galtonian Perspective on Shrinkage Estimators. *Statistical Science*, *5*, 147-155.

'Student' (William Gossett). The Probable Error of a Mean. *Biometrika, 6*, 1-25.

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.

Taylor, B.L. & Gerrodette, T. (1993). The uses of statistical power in conservation biology: The vaquita and Northern Spotted Owl. *Conservation Biology, 7*, 489-500.

Thomas, L. & Juanes, F. (1996). The importance of statistical power analysis: an example from Animal Behaviour. *Animal Behaviour, 52*, 856-859.

Thomas, L. & Krebs, C.J. (1997). A review of statistical power software. *Bulletin of the Ecological Society of America, 78*, 126-139.

Thompson, B. (in press). *Foundations of Behavioral Statistics: An insight-based approach.* New York: Guilford.

Thompson, B. (2002a). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, *80*, 64-71.

Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 24-31.

Thompson, B. (2000). A suggested revision to the forthcoming 5th edition of the APA *Publication Manual*. Retrieved February 14, 2002, from http://www.coe.tamu.edu/~bthompson/ apaeffec.htm

Thompson, B. (1999a). If statistical tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 165-181.

Thompson, B. (1999b). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*, 191-196.

Thompson, B. (1999c). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*, *11*, 157-169.

Thompson, B. (1998a). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*, 33-38.

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. *American Psychologist*, *53*, 799-800.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher, 25*, 26-30.

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

Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. *Journal of Experimental Education, 66*, 75-83.

Toft, C.A. & Shea., P.J. (1983). Detecting community-wide patterns: estimating power strengthens statistical inference. *American Naturalist, 122*, 618-625.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

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

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003). Improving precision and reducing bias in biological surveys: Estimating false negative error rates. *Ecological Application, 13*, 1790–1801.

Tyron, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

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.

Van Winkle, W., Vaughan, D.S., Barnthouse, L.W. & Kirk, B.L. (1981). An analysis of the ability to detect reductions in year-class strength of the Hudson River white perch (*Morone americana*) population. *Canadian Journal of Fisheries and Aquatic Science, 38*, 627-632.

Vincete, K.J. (1997). Four reasons why the science of psychology is still in trouble, *Behavioral and Brain Sciences, 21*, 224-225.

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

Wagenmakers, E.J. & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192-196.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4*, 212-213.

Ware, J.H., Mosteller, F. & Ingelfinger, J.A. (1986). P values. In J.C. Bailar & F.Mosteller (Eds.). *Medical Uses of Statistics*. Waltham, MA, USA: NEJM Books.

Weinbach, R.W. (1989). When is statistical significance meaningful? A practice perspective. *Journal of Sociology and Social Welfare, 16*, 31-37.

Weisburd, D., & Lum, C.M., & Yang, S.M. (2003). When can we conclude that treatments or programs "don't work"? *The Annals of the American Academy of Political and Social Science, 587*, 31-48.

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.

Wilson, W.R. & Miller, H. (1964). A note on the inconclusiveness of accepting the null hypothesis. *Psychological Review*, *71*, 238-242.

Wilson, W.R., Miller, H. & Lower, J.S. (1967). Much ado about the null hypothesis. *Psychological Bulletin, 67*,188-197.

Wintle, B.A., McCarthy, M.A., Parris, K.P. & Burgman, M.A. (2004). Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications, 14*, 703-712.

Wintle, B.A., McCarthy, M.A., Volinsky, C.T. & Kavanagh, R.P. (2003). The use of Bayesian model averaging to better represent the uncertainty in ecological models. *Conservation Biology, 17,* 1579-1590.

Wolfe, R., & Hanley, J. (2002). If we're so different why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal, 166*, 65-66.

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

Wulff, H.R., Andersen, B., Brandenhoff, P. & Guttler, F. (1987). What do doctors know about statistics? *Statistics in Medicine*, 6, 3-10.

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

Ziliak, S. & McCloskey, D. (2004). Size matters: The standard error of regressions in the *American Economic Review*, *Journal of Socioeconomics, 33*, 665-675.

Appendix. Examiner's Reports

EXAMINER 1: Geoff Loftus

My comments will be brief because my evaluation is straightforward: This dissertation is one of the best written and organized, most comprehensive, and most scientifically important documents on the topic of statistical methodology—or indeed on any scientific topic—that I have ever read. I can only hope that Ms.Fidler finds an outlet for it that will allow it to enjoy widespread readership. I myself will plan to organize a graduate seminar during 2006 whose sole purpose will be to go through the dissertation. I wish every researcher who uses statistics would read it. I believe that it could easily be published as is, or with very minor editorial changes. This is not the kind of praise that I bestow lightly.

For many years numerous branches of experimental science have been beset with sufficient statistical noise that some form of statistical analysis is necessary and routine. Several of these sciences have come to rely on the form of null hypothesis significance testing (NHST) that is the central focus of Fidler's dissertation. My own discipline, psychology, has been such a science. It has always been clear to me and a to a relatively small cadre of others in my field that such reliance is, for many reasons, highly inimical to understanding data sets in particular, and to scientific progress in general. As Fidler points out, numerous articles, appearing in prestigious journals, have appeared over the years to point out this difficulty; and yet, at least in psychology, the arguments have had little to no effect on scientific progress.

I have known about these issues and the attendant articles because they have appeared in my own discipline. Meanwhile, I have been dimly aware of similar issues in other disciplines, such as medicine. However, prior to reading Fidler's dissertation, I did not have an appreciation of (1) the scope of the difficulties in other disciplines or (2) the results of efforts to address these difficulties in other disciplines[38]. I believe that describing the issue as it has arisen and has been addressed in three separate disciplines is a major strength of the dissertation.

Other strengths include a fairly comprehensive approach to the problem as a whole. Fidler very nicely describes and investigates the following interrelated factors:

1. The philosophical and logic basis of NHST.
2. The problems with NHST.

---

[38] One caveat is appropriate here: As I was reading Fidler's dissertation I received, purely by coincidence, a request to review a draft of a book, submitted to a major publisher, written by two economists that addressed the same topic material. Although the authors of this book covered much of the same ground and reached many of the same conclusions as Fidler did, the scholarship was so poor, the organization so scattered, and writing so generally awful that I reluctantly advised that the manuscript never see the light of day.

3. Solutions to these problems.
4. Potential new and/or unexpected problems with the solutions.
5. How the foregoing four factors play out in three separate disciplines.

Moreover, Fidler includes not only discussion of the problems (which, in my mind would have sufficed for a dissertation) but also describes a fairly broad set of data that she has collected. These data involve both assessing the use of various statistical techniques over the years in existing journal, and a set of actual psychological experiments designed to accomplish goals such as testing reactions to and understanding of the logic both of NHST and of the favored alternative (confidence intervals).

Finally the quality of writing and of organization was refreshingly clear and to the point. That is in contrast to many academic documents whose authors adopt a kind of dry, jargon-laden style that appears to be designed more to impress other academics than to actually convey important information.

This completes my comments. They are brief, but with a dissertation whose quality is as high as this one, brevity is appropriate.

EXAMINER 2 : Frank Schmidt

My overall evaluation is that this is an excellent dissertation. It examines in thorough and analytic detail a critical, widespread, and puzzling deficiency in data analysis and scientific research. It illuminates nicely the history and psychology underlying the failure of researchers in most areas to change their data analysis practices in the face of overwhelming logical arguments against traditional practice. It was a pleasure to read such a well thought out exposition of the perplexing problem.

After being tightened up and shortened to eliminate redundancy and repetition, this dissertation should be published as a book, in my judgement. It should be published by a publisher, such as the American Psychological Association, that will advertise and promote the book appropriately, so as to maximize its impact on practice. Studies of change processes show that in many cases change is negligible for an extended period of time and then is very rapid after a “tipping point” is reached. This dissertation in the form of a book could be the stimulus of the tipping point in the reform of data analysis and interpretation practice.

To speed this process up, I recommend that the author circulate the completed dissertation electronically to key individuals, including all those listed at the beginning of the text as having been interviewed by the author and perhaps all those cited in the dissertation. It is important that this work be disseminated as widely as possible as quickly as possible. (Disseminating the dissertation electronically in this manner will not prevent of hinder its publication as a book.)