

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology, 12*, 825-853.

© Sage Publications

Journal website: <http://www.sagepub.com/journal.aspx?pid=147>

This article may not exactly replicate the final version published in the journal. It is not the copy of record."

Past and future APA guidelines for statistical practice

Sue Finch¹, Neil Thomason² & Geoff Cumming¹

1 = La Trobe University, Melbourne, Australia

2 = The University of Melbourne, Melbourne, Australia

Abstract

We review the publication guidelines of the American Psychological Association (APA) since 1929 and document their advice for authors about statistical practice. Although the advice has been extended with each revision of the guidelines, it has largely focussed on Null Hypothesis Significance Testing (NHST) to the exclusion of other statistical methods. In parallel, we review over 40 years of critiques of NHST in psychology. Until now, the critiques have had little impact on the APA guidelines. The guidelines are influential in broadly shaping statistical practice, although in some cases recommended reporting practices are not closely followed. The guidelines have an important role to play in reform of statistical practice in psychology. Following the report of the APA's Task Force on Statistical Inference, we propose that future revisions of the guidelines reflect a broader philosophy of analysis and inference, provide detailed statistical requirements for reporting research, and directly address concerns about NHST. In addition the APA needs to develop ways to ensure that its editors succeed in their leadership role in achieving essential reform.

Past and future APA guidelines for statistical practice

In this paper we will recount and compare two histories, and draw implications for the role of future APA publication guidelines. The first history describes the changing standards for reporting statistical inference found in the publication guidelines of the American Psychological Association (APA). The guidelines have served as the reference standard for authors and editors for over 70 years. The second history is of the ongoing critique of NHST use in psychology and the introduction of alternative approaches to data analysis. What is remarkable is how little these two histories overlap. The APA guidelines have failed to reflect the compelling arguments made against the routine use of NHST; they have given little indication that there are alternative methods of analysis. They have failed to warn authors of pitfalls in the use of NHST, and of damage NHST practices appear to have caused.

It is important to reflect on these histories now. The APA's Task Force on Statistical Inference (TFSI) has published proposed recommendations (Wilkinson & Task Force on Statistical Inference, 1999, henceforth TFSI report) for changes, only a few of which have been included in the fifth edition of the APA Publication Manual (APA, 2001). The report canvasses issues and problems in statistical analysis more broadly than any APA Manual. It puts forward a wide range of proposals, three of which we regard as especially important: First and most pervasively, careful thinking and substantive interpretation must be the main focus.

On this issue the TFSI first quoted Fisher (1935):

Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do the authors of text-books written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proved value in their own subject. (p. 49)

The TFSI returned to the issue of substantive interpretation to conclude their paper:

More than 50 years ago, Hotelling, Bartky, Deming, Friedman, and Hoel (1948) wrote, "Unfortunately, too many people like to do their statistical work as they say their prayers—merely substitute in a formula found in a highly respected book written a long time ago" (p. 103). Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. If editors keep in

mind Fisher's (1935) words [quoted above]... then there is less danger of methodology substituting for thought. Statistical methods should guide and discipline our thinking but should not determine it. (p. 603)

The second TFSI proposal we wish to highlight was for a change in philosophy of analysis. Rather than recommending particular methods of analysis, the TFSI suggests using Occam's razor in choosing "a minimally sufficient analysis" (p. 598):

The enormous variety of modern quantitative methods leaves researchers with the nontrivial task of matching analysis and design to the research question. 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. Do not choose an analytic method to impress your readers or to deflect criticism. If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. (p. 598)

Third, the TFSI proposed a move away from routine reliance on NHST as a primary means of analysing data to exploring, summarising and analysing data using visual representations, effect size measures and confidence intervals, amongst other things. Despite some pressure, the TFSI did not choose to ban NHST.

We take the position that reform like that recommended by the TFSI is highly desirable and urgent, and that NHST, especially as practised by psychologists, has severe problems. In this paper we do *not* attempt to review and evaluate all the arguments about the logical basis of NHST and its role in psychology. We recognise that there are two main foci of the debate about NHST: (i) its inherent logic and what it can and cannot do, and (ii) the way it has been practised and interpreted by psychologists. Debate on both continues, as well as on what the consequences for practice should be. Our history attempts to identify many of the key papers, while not attempting to cover the full debate in detail.

There are salutary lessons in the histories we present. 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. In light of these lessons we consider how APA can best support the far-reaching changes recommended by the TFSI.

Two histories to 1994: Statistical reform, and APA guidelines

Before 1974. Over the past 70 years, the APA has published a number of documents describing publishing standards for articles appearing in its journals. The 1929 Instructions in regard to preparation of manuscripts, published in the Psychological Bulletin, focused on details of style and layout without specifying how results should be reported (“Instructions”, 1929). Similarly, the 1944 guidelines contained no explicit instructions about how to report study results (Anderson & Valentine, 1944).

In the first edition of the APA Publication Manual, published in 1952, the following guidelines appeared: “The section on results should give enough data to justify the conclusions. Special attention should be given to tests of statistical significance and to the logic of inference and generalization from empirical observations” (APA, 1952, p. 397). Further, in describing economy in presentation of tables, it was stated that: “Extensive tables of nonsignificant results are seldom required. For example, if only 2 of 20 correlations are significantly different from zero, the 2 significant correlations may be mentioned in the text, and the rest dismissed with a few words” (APA, p. 414).

In revisions to the first edition (APA, 1957, 1967), some further guidelines were provided in relation to tests of statistical significance. First, authors were cautioned “against inferring trends from data which fail by a small margin to meet the level of significance adopted. Such results are most economically interpreted as a function of chance and should be reported as nonsignificant” (APA, 1967, p. 13). Note that this caution encourages interpreting statistically non-significant results as due to chance—a common misconception.

A further addition in the revisions (APA, 1957, 1967) referred to the special case in which a paper reports the predictive validity of methods of measurement:

the results must always show the degree of relationship between the measure and the criterion, and its practical value. The relationship should be reported in such terms as... [for example] correlation ... It is not sufficient merely to show that the relationship is nonchance in terms of the level of significance. (APA, 1967, p. 13, emphasis in original)

This is an early Manual reference to the need, at least in one particular situation, to report effect size and practical importance.

The 1967 revision (APA, 1967) introduced the requirement for an abstract and stated that “every abstract should contain at least the trend or direction of results. ... and the

significance levels of results also should be included” (p. 12). It is unclear whether the statistical significance levels referred to were exact or relative p -values, or a priori alphas.

For over 20 years, these statements were the only APA guidelines that psychologists had for preparing statistical evidence for publication in scholarly articles. Fortunately, many important papers about use of statistics were appearing in the psychological literature.

In the period between the first (APA, 1952) and second (APA, 1974) editions of the Manual, criticism of standard NHST use in the social sciences was growing. In 1969, American Psychologist published John Tukey’s invited address to the APA. Tukey emphasized exploratory data analysis (EDA) rather than only ‘sanctified’ confirmatory analysis such as NHST. Tukey quoted an anonymous colleague who referred to the “stultifying uniformity of statistical usage” (p. 90) and argued that:

Following a rule book for research seems to stimulate the attack on trivial problems. The great challenge is to teach investigators to formulate questions that have a chance of leading somewhere, not to be too tightly bound in the formulation by a preconceived model of research design. Only after the formulation ... need there be attention to the procedures to be adopted for collecting and evaluating evidence—not right away changing the questions to fit a standard procedure. (p. 90; emphasis in original)

Morrison and Henkel’s (1970) anthology included reprints of 30 articles written by sociologists, psychologists and statisticians, originally published between 1941 and 1969. The articles were largely critical of NHST; none defended standard NHST practice in psychology. In their preface, Morrison and Henkel stated:

In the behavioral sciences in general the overwhelming practice by both researchers and those responsible for statistical training has been to ignore the issues raised by the critics and to continue doing things as before. Thus the preponderance of the negative side of the “debate” in this volume does not represent so much bias as redress, since the amount of behavioral science writing that implicitly supports the tests is far greater than that which is critical. (p. x)

Important discussions and critiques of NHST were published in Psychological Bulletin (e.g. Bakan, 1966; Edwards, 1965; LaForge, 1967; Lykken, 1968; Rozeboom, 1960), Psychological Review (e.g. Binder, 1963; Grant, 1962; Wilson & Miller, 1964) and

elsewhere (Meehl, 1967; Nunnally, 1960). Meehl, for example, noted the important difference between a substantive theory and a statistical hypothesis derived from that theory, and the unfortunate practice of interpreting the result of the statistical test as also directly quantifying support for the substantive theory. More appropriate alternatives to NHST recommended included measures of strength of association (e.g. Nunnally, 1960), confidence intervals (e.g. Grant, 1962; LaForge, 1967) and Bayesian methods (e.g. Bakan, 1966; Edwards, 1965; Edwards, Lindman & Savage, 1963). We know of no defenses of NHST published by psychologists in this period.

Psychological Bulletin published in 1957 a short article by Chandler that clarified the difference between NHST and confidence intervals. Chandler also discussed the importance of statistical power:

Although texts in psychological statistics do not seem to place a great deal of emphasis upon the power of a test, power is the basic concept responsible for one's employing statistical tests as a basis for taking action on an H [Hypothesis]. If this were not so, to test an H at the 5 percent level of significance, one could simply draw from a box of 100 beads—95 white and 5 red—a bead at random and adopt the convention that he would reject the H whenever a red bead appeared. (p. 430)

Jacob Cohen's (1962, 1965, 1969) important works strongly advocated statistical power. Cohen's (1962) paper, published in the Journal of Abnormal and Social Psychology, documented the alarmingly low power of abnormal-social psychological research. Cohen's (1969) book explained how to calculate power for a range of experimental designs and provided tables to help with the calculations.

Psychology journals also published empirical evidence about the difficulties psychologists had understanding NHST (e.g. Beauchamp & May, 1964; Rosenthal & Gaito, 1963, 1964). In 1971, Tversky and Kahneman's much-cited paper on the law of small numbers misconception appeared. It warned psychologists that their over-reliance on small samples for testing hypotheses is problematic, and underlies widespread misunderstanding of replication and power. This influential paper helped spawn an enormous research enterprise that continues to investigate the psychology of judgement under uncertainty. However, these findings and other issues raised about NHST made no detectable impact on the 1974 edition of the APA Manual.

The 1974 Manual. In the foreword to the 1974 Manual, Arthur Melton described the intent of the Manual:

This 1974 edition of the Manual maintains the original intent of the early guides to help authors with the endless detail of manuscript preparation. However, this edition broadens that intent in several ways. It recognizes that what were once suggestions to authors of psychology articles are now course content for students of psychology. It also recognizes that, in 1929, APA could gently advise its authors on style because there were only 200 or so who reached print in the 4 APA journals. Today, ... [APA] editors ... consider more than 6,000 manuscripts a year ... Without APA style conventions ... clear communication would be jeopardized. (APA, 1974, p. 5)

The Introduction also commented on the broadening of intent:

In 1967, when the Publication Manual was revised, it was intended exclusively for APA authors. ... the second edition is published for a more varied audience, including graduate and undergraduate students, editors, publishers, authors of non-APA journals, copy editors, typists and printers. For this wider readership ... Chapter 1 ... discusses ... how to conceptualize and structure ideas and data into elements of an article ... (p. 7)

Some details for reporting statistical tests were made explicit in the 1974 Manual (APA, 1974): “In reporting tests of significance ... include information concerning the obtained magnitude or value of the test, the degrees of freedom, the probability level, and the direction of the effect” (p. 18). In a later section, there are examples showing how to report inferential statistics such as: “As predicted, the first-grade girls reported a significantly greater liking for school than did the first-grade boys, $t(22) = 2.62, p < .01$ ” (p. 39). The examples indicate that the term ‘probability level’ referred to the (relative) p -value and the term ‘magnitude’ referred to the value of the test statistic (e.g. $t = 2.62$), rather than to an effect size measure.

In the example above, there was no information about the means and standard deviations of the groups, or about the difference between the groups. There was no information about how much the girls and boys liked school or how much they differed in their liking of school. Further, in the section on results authors were simply told to “summarize the collected data” (APA, 1974, p. 18). There was no explicit requirement to

include, for example, measures of central tendency or of spread. Indeed, the section on results focused more on the choice between tables and graphs than on the type of descriptive statistics to be reported. As in the 1967 revision, authors were advised to include in their abstract a results summary: “Summarize the data or findings, including statistical significance levels, if any, as appropriate” (APA, p. 15).

The following caution, a modification from earlier revisions (APA, 1957, 1967), appeared in the section on results:

Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return: Take what’s coming to you, but no more. (APA, 1974, p. 19)

This caution mandates using the “usual” significance level to make dichotomous decisions. Like the earlier versions, it actively encourages misinterpreting statistically non-significant results as due to chance—a common misconception. The 1957 and 1967 revisions of the 1952 Manual had also required reporting effect size and practical importance, along with NHST results, in the special case of presenting information about the predictive validity of measures; these points were not included in the 1974 Manual.

The 1974 Manual included a relatively long discussion of the “perennial and vexing problem” of “negative results”—that is, statistically non-significant results (APA, 1974, p. 21). Situations when editors might be interested in negative results were described. If a theory predicts that a difference or correlation should be found, a result to the contrary may be of interest; however, the “burden of methodological precision falls heavily on the investigator who reports negative results. The greater the corpus of methodologically sound results supporting the theory, the heavier the burden” (p. 21). Negative results that illustrate methodological weaknesses in other research “are also valuable” (p. 21). However, “Failure to replicate results of a previous investigator, using the same method but a different sample, is generally of questionable value. A single failure may merely testify to sampling error or to the conclusion that one of the two samples had unique characteristics” (p. 21).

This concern and discussion of negative results may reflect the debates about the inappropriateness of equating a theory with the null hypothesis, and about potential biases for and against the null hypothesis (e.g. Binder 1963; Edwards, 1965; Wilson, Miller & Lower, 1967). However, it is striking that the Manual’s discussion of negative results makes no

mention of the role of sample size or of statistical power. Psychologists relying on the 1974 APA Manual as a guide to statistical practice would remain unaware of the debate about NHST, alternative ways of analyzing data, Cohen's (1962) call to calculate power, and Tversky and Kahneman's (1971) important findings. Guided by the APA Manual, many psychologists in good faith continued their seriously flawed statistical practices.

1974 to 1983. In the period between the publication of the second and third editions of the APA Publication Manual (1974, 1983), articles condemning the use of NHST continued to appear. Notable contributions from this time included Paul Meehl's (1978) article on the problems of scientific progress in psychology, and Lee J. Cronbach's (1975) call to "exorcise the null hypothesis" (p. 124). Cronbach suggested focussing on descriptive information and confidence intervals, and warned against using statistical significance as a criterion for deciding which results to report.

Carver (1978) made a much-cited contribution that argued that the way researchers, especially in education, use the NHST significance criterion represents "a corrupt form of scientific method" (p. 378). He recommended that researchers wishing to use NHST report effect sizes and statistical power, and also stated that "standards errors and confidence intervals should not be ignored" (p. 396).

Researchers continued to empirically demonstrate people's difficulties in understanding statistical tools and principles. By 1983, Tversky and Kahneman's (1971) paper had been cited nearly 200 times in the Social Sciences Citation Index. New tools for data description, exploration and integration were further developed. For example, in 1977 Tukey's seminal book on EDA was published. As far as we know, no explicit defenses of standard NHST practices were published in the psychological literature during this period. Of course, the universality of these practices, the APA Manual and the uncritical presentation of NHST in textbooks no doubt were taken as endorsements.

In 1976 Gene Glass, in his presidential address to the Annual Meeting of the American Educational Research Association, introduced a method of statistically integrating the results of many studies—a method he called meta-analysis. Glass' method avoided difficulties of narrative reviews that often relied on counts of statistically significant and statistically non-significant results (e.g. Glass, 1976; Hunter & Schmidt, 1990). Two important books on meta-analysis appeared before 1983 (Glass, McGaw & Smith, 1981; Hunter, Schmidt & Jackson, 1982).

In this address, Glass (1976) emphasized the effect size of a therapeutic intervention: the mean difference in the outcome variable between treated and untreated subjects divided by

the within-group standard deviation. The application of meta-analysis in education and psychology raised the issue of how to quantitatively compare outcome measures that were obtained using different measurement units. Cohen (1969) had also proposed this kind of standardized effect size measure (Cohen's d) for use in power analysis.

The use of raw and standardized differences, and measures of association as indices of the magnitude of study effects, was discussed in a range of psychological journals between 1974 and 1983 (e.g. Carter, 1979; Carver, 1978; Cohen, 1973; Craig, Eison & Metze, 1976; Kern & Lewis, 1979; Maxwell, Camp & Avery, 1981; O'Grady, 1982). The measures discussed were not all new – Hays (1963), for example, had described omega-squared in his popular textbook, and Friedman had provided a guide for estimating a correlation measure of effect size in a Psychological Bulletin article in 1968.

The 1983 Manual. The 1983 Manual echoed the comments on the intent of the 1974 Manual. Remarkably, in the 1983 Manual recommendations for statistical presentation were essentially unchanged. The examples of reports of statistical inference in text were reproduced from the 1974 Manual with the addition of the values of the sample means (APA, 1983, p. 81). The need to report descriptive statistics (means and standard deviations) was also made explicit in the section on writing up results (p. 27).

The statement on negative results was that “Negative results should be accepted as such without an undue attempt to explain them away” (APA, 1983, p. 28). This somewhat cryptic recommendation was unelaborated. The caution indicating that statistically non-significant results should be interpreted as due to chance was dropped.

Overall, in the 1983 Manual the emphasis remained on how to write clearly, editorial style and procedural details for writing and submitting manuscripts. Although the Manual aimed to give guidance on content, scant attention was given to presenting or interpreting data and NHST results. Terms such as ‘probability level’ and ‘significance level’ were not clearly defined; this probably reflected the confusion about NHST concepts claimed by critics to be widespread among psychologists. There was no mention of EDA, effect sizes, confidence intervals, statistical power, meta-analysis, or plausible misinterpretations of NHST.

1983 to 1994. Clearly the reform message had not been heard, and so discipline leaders reiterated their messages about alternative methods of data analysis and the problems of NHST (e.g. Carver, 1993; Cohen, 1990, 1994; Loftus, 1993a; Meehl, 1990; Oakes, 1986; Rosnow & Rosenthal, 1989; Schmidt, 1992). As incoming editor of Memory and Cognition, Loftus (1993b) made a laudable attempt to discourage reliance on NHST in favor of visual displays of means and standard errors; Finch, Cumming, Williams et al. (2001) studied the

effectiveness of Loftus' reform. G. R. Loftus (personal communication, 6 April, 2000) reported however that his "recommendations were not met at the time with the kind of widespread enthusiasm and compliance that I had ...anticipated".

Researchers continued to document widespread misinterpretations of NHST made by psychologists and textbook writers (e.g. Gigerenzer, 1993; Oakes, 1986). Many texts on meta-analysis appeared, and many meta-analyses were published in education and psychology journals. The practical application of meta-analysis raised many issues. Hunter and Schmidt's (1990) book, for example, discussed analytic methods that they claimed assessed a variety of potential threats to the valid combination of studies. Methods of meta-analysis remain to the present under development and debate.

Cohen (1988) published a revised edition of his book on power. Sedlmeier and Gigerenzer (1989) replicated Cohen's (1962) study of statistical power, 24 years later, and reported a "considerable decrease of power" (p. 313). Clearly, Cohen's (1962) study and enthusiastic advocacy of power had made little impact on research or editorial practice. Sedlmeier and Gigerenzer reported that they found "almost no concern with power" (p. 313).

The 1994 Manual. More substantial changes appeared in the fourth edition of the Manual (APA, 1994). Presentation of statistics was one aspect that the revision task force was authorized to change (APA, p. xxii). For example, ways of reporting a range of different inferential test results were described. Importantly, the need to consider statistical power was discussed for the first time:

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)

The reporting of negative results without a power analysis was mentioned as one of five defects found by editors in submitted papers (p. 3).

A recommendation to specify the a priori significance level appeared for the first time, and the difference between alpha and *p*-values was described:

One [type of probability] refers to the a priori probability you have selected as an acceptable level of falsely rejecting a given null hypothesis. This probability [is] called the alpha level ... The other kind of probability refers to the a posteriori likelihood of obtaining a result that is as extreme as or more extreme than the actual value of the statistic you obtained, assuming that the null hypothesis is true. (p. 17; emphasis in original)

The Manual recommended that “you should routinely state the particular alpha level you selected for the statistical tests you conducted...” but then continued “If you do not make a general statement about the alpha level, specify the alpha level when reporting each result” (p. 17). It then defined and explained the exact p-value and stated “You can report this distinct piece of information [exact p-value] in addition to specifying whether you rejected or failed to reject the null hypothesis using the specified alpha level” (p. 17). Finally it permitted the use of relative p-values: “If you do not wish to report the exact probability, you can report the commonly used probability value that is nearest to it...” (p. 18). Overall, the p-value was presented as an adjunct to a dichotomous decision based on an a priori alpha.

The explanation of alpha and p-values may have assisted researchers in their NHST practices, but may if anything have entrenched these further. It was hardly a contribution to reform.

Expanded treatment of NHST did not extend to warning researchers about misuse of the ambiguous term ‘significant’ (Boring 1919; Carver, 1993). In some examples of how results should be reported (APA, 1994, pp. 17-18) the correct term ‘statistically significant’ was carefully used, but in others (p. 113) ‘significant’ was used without this qualification. Indeed at the very start, when discussing assessment of research quality, the Manual inadvertently illustrated the ambiguity when it urged researchers to ask “Is the research question significant...” (p. 3).

For the first time, general reporting of effect size information was encouraged:

Neither of the two types of probability values 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. (APA, 1994, p. 18)

The admirable advice to report effect size measures was thus compromised in two ways. First, the importance and magnitude of an effect were conflated. Second, the qualification that effect size measures may usually be calculated from commonly reported information undermines the recommendation to report effect size measures, and fails to recognize that a central reason for reporting them is to inform substantive, practical interpretation of results. Careful examination of effect sizes is necessary but not sufficient for such interpretation.

The 1994 Manual retained a strong emphasis on NHST, and space was devoted to explaining NHST concepts. Authors were still advised to describe “the findings, including statistical significance levels” (p. 10) in writing their abstracts. The examples of statistical text, tables and figures predominantly referred to NHST. There was no mention of EDA or explicitly of confidence intervals. Standard errors only appeared in examples of tables and figures, and then only occasionally. Meta-analysis was mentioned only in the section on preparing the reference list. There was no indication that NHST is widely and seriously misunderstood by researchers.

The role of the APA Manual

APA Manuals have always had the stated intention of supporting clear communication of psychological research. They have provided advice on good writing and good research practice, but most space by far has been devoted to extensive detail about the style required for publication.

Our review of 70 years of APA publication guidelines and four editions of the APA Manual reveals that, until recently, the guidelines provided little reason for psychologists to change their analytic and reporting practices.

On the whole, APA Manuals have provided sparse details about methods of analysis and interpretation. The guidelines have failed to acknowledge the debates, concerns and relevant empirical findings of the day. They have failed even to acknowledge that concerns have been voiced over many decades. In the fourth edition, the first few clear recommendations to changes in statistical reporting practice appeared (APA, 1994), but these were recommendations, rather than requirements, and covered only some of the important issues.

Although the focus of the publication guidelines is on how to put an article together in the right format and style, they also make general statements on good research practice. Beyond shaping presentation the guidelines have taken on a de facto role as arbitrating appropriate statistical technique by including details about how to report just certain statistical

procedures. Some researchers may unfortunately but naturally merely mimic the simple examples that they find in the Manual—and these are the very examples of reporting style they see published. These examples assume NHST and neither suggest nor illustrate alternatives to it. The failing of the Manuals is the way NHST is presented as the method of analysis.

There is a further important aspect to the role of the APA Manual: Authors often do not follow Manual guidelines (Vacha-Haase & Ness, 1999; Vacha-Haase & Nilsson, 1998) and, in relation to statistical procedures, editors, associate editors and manuscript reviewers have enforced the guidelines only selectively. For example, authors rarely include information about a priori alpha levels, and only infrequently is there any sign that power is taken seriously (Finch, Cumming & Thomason, 2001). They rarely mention *p*-values in their abstracts, despite a recommendation since 1967 to do this. Vacha-Haase, Nilsson, Reetz, Lance and Thompson (2000) reviewed 10 empirical studies of reporting of effect size measures (taken to include, for example, *r* and R^2) in 23 journals. With the exception of the Journal of Applied Psychology, “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’” (Vacha-Haase et al., p. 419, emphasis in original).

Failure to follow certain guidelines may be partly due to the mixture of mandates and ‘softer’ recommendations in the Manuals. Soft recommendations about appropriate statistical practice stand in stark contrast to the mandates about formatting and style. On the APA’s (1994) encouragement to report effect sizes, Thompson (1999) stated: “To present an ‘encouragement’ 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). Many statements on style are interpreted by authors, editors and copy editors as mandates, although it is noted in the foreword: “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” (APA, 1994, p. xxiii).

Over several decades the official APA guidelines largely failed to address the highly desirable changes advocated by reformers. The guidelines presented a limited view of statistical practice by supporting NHST as the standard for analysis with little indication that it is problematic. At the same time, particular APA guidelines have not been followed consistently. The slow response to the few reform recommendations that did appear in the

1994 Manual suggests that such recommendations by themselves are not a potent way to modify researchers' publishing practices. These conclusions have important implications for the changes proposed by the TFSI. We consider these implications, and policy consequences, after a brief review of recent discussions of reform and the fifth edition of the Manual (APA, 2001).

Debate about NHST since 1994

Recently, the critique and debate about NHST has intensified (e.g. Falk & Greenbaum, 1995; Grayson, Pattison & Robins, 1997; Hammond, 1996; Kirk, 1996; Loftus, 1996; Schmidt, 1996; Thompson, 1996). In 1997, a special issue of Psychological Science debated the issues (see for example, Shrout, 1997). Harlow, Mulaik and Steiger's (1997) anthology What if there were no significance tests? included contributions from key commentators including Abelson, Hunter, Meehl, Rozeboom and Schmidt. In summarizing the contents, Harlow (1997) concluded that all contributors agreed that psychologists should focus on confidence intervals and routinely examine effect size measures and statistical power. Additionally, most authors supported the testing of specific (non-zero) hypotheses: tests of risky but specific hypotheses have the potential for strong theory corroboration. In contrast, "the overriding view ... is that NHST [typically with a zero null hypothesis] may be overused and unproductive, particularly when used as simply a dichotomous decision rule" (Harlow, 1997, p. 12). Certain limited uses of NHST were defended (e.g. Abelson, 1997; Mulaik, Raju & Harshman, 1997).

Few defenses of NHST were published in the psychological literature prior to 1994 (e.g. Chow, 1988), but a number have appeared in the past 5 years (Abelson, 1997; Chow, 1996; Cortina & Dunlap, 1997; Frick, 1995; 1996; Hagen, 1997; Harris, 1997; Mulaik et al., 1997). They generally agreed that standard practices are flawed in some ways, but they argued that NHST has some function in certain situations. The most enthusiastic defender was Chow whose synopsis of his 1996 book, Statistical significance: Rationale, validity and utility, appeared in 1998 in Behavioral and Brain Sciences. Chow (1996, 1998) defended NHST as a mechanism for evaluating hypotheses rather than theories and argued that it was important to distinguish hypotheses at various levels—statistical, experimental, research, and substantive. He argued that "Statistical significance only means that chance influences can be excluded as an explanation of data" (Chow, 1998, p. 169). He argued that effect size measures and confidence intervals could not fulfill this role, and that critics of NHST overplayed the role of these alternative statistics. Chow also questioned the validity of power analysis and meta-

analysis. However even Chow (1996) held that NHST "plays only a very limited ... role in empirical research" (p. x). He also agreed that standard practice is in need of improvement.

Task Force on Statistical Inference

In 1996, the APA responded to the debate by forming the Board of Scientific Affairs Task Force on Statistical Inference. Members of, and advisors to, the TFSI included many advocates of reform, some of whom such as Abelson defend the use of NHST in certain contexts. The TFSI's 1999 report followed a recommendation of the APA's Board of Scientific Affairs that "before the TFSI undertook a revision of the APA Publication Manual, it might want to consider publishing an article in American Psychologist, as a way to initiate discussion in the field about changes in current practices of data analysis and reporting" (TFSI report, p. 594).

The TFSI report made recommendations covering many aspects of research design, method, analysis and interpretation. It discussed assignment of participants to experimental conditions, measurement, EDA, interpreting computer output and many other aspects of statistical analysis and interpretation. It provided references to many important papers dealing with a wide range of issues in the application of methods of analysis. In addition the report explicitly recognized the role of the APA Manual to go beyond reporting style by setting out principles of good statistical practice.

As we noted earlier, the TFSI report emphasized the need to choose "a minimally sufficient analysis" (p. 598) and the central importance of substantive interpretation of results. It warned against the routine use of any particular method, although it was cautious in its recommendation about NHST:

Some had hoped that this task force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought that there were enough counterexamples (e.g., Abelson, 1997) to justify forbearance. Furthermore, the committee believed that the problems raised in its charge went beyond the simple question of whether to ban significance tests. (pp. 602-603)

Nonetheless, the TFSI's message about NHST was a marked shift from all the APA Manuals: "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" (TFSI report, p. 599). The TFSI recommended reporting effect sizes and confidence intervals for the principal

outcomes of a study. The need to think about results in the context of past and future research was emphasized:

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. (TFSI report, p. 599).

In sum, the TFSI report recognized the need for psychologists to broaden their repertoire of statistical tools, and that choice of tools is dependent upon the problem at hand. It emphasized that researchers should primarily focus on substantive interpretation. The report should be recognized as an immense step forward in the urgent and essential reform process, both for the broad official agenda it sets for consideration and for the impetus it gives towards concrete policy measures to effect widespread change.

The fifth edition of the APA Manual (2001)

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.

There is some good news. On two important reform issues there are strong statements. The paragraphs on effect size are strengthened from "You are encouraged to provide..." (APA, 1994, p. 18) to:

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. You can estimate the magnitude of effect or the strength of the relationship with a number of common effect size estimates... The general principle to be followed... is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (APA, 2001, pp. 25-26)

In addition the “failure to report effect sizes” (p. 5) is now listed as a defect in a paper. An example ANOVA table (p. 162) now includes a column showing effect sizes.

Second, there is for the first time a paragraph on confidence intervals:

The reporting of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended. (APA, 2001, p. 22)

We hope that reform-minded editors, referees, and authors will be able to use these statements to good effect to support their efforts.

However the main impression that the fifth edition gives, from a statistics perspective, is that it is only a modest revision of the 1994 edition. In the introduction it is stated “The statistics section has been largely rewritten to reflect emerging standards in the field (although there are still a number of disagreements on presentation)” (p. xxiv). This seems to us to substantially overstate the amount of revision, and to trivialise the reform debate. Many paragraphs remain exactly or almost exactly as before, including the paragraph on statistical power, and the statement that the abstract “should describe...the findings, including statistical significance levels...” (APA, 1994, p. 10; 2001, p. 14). That recommendation, which is rarely followed (Finch, Cumming, & Thomason, 2001), seemed anachronistic in 1994; its perpetuation in the fifth edition raises questions as to how thoroughly all the statistical material was scrutinised during preparation of this latest edition.

There is for the first time recognition of the NHST debate, in a carefully non-committal way:

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*. These issues include, but are not limited to, the reporting and interpretation of results of hypothesis tests, the selection of effect size indicators, the role of hypothesis-generating versus hypothesis-testing studies, and the relative merits of multiple degree-of-freedom tests. A discussion of these and other issues can be found in Wilkerson [sic] (1999). (APA, 2001, p. 21)

That is the only reference to the TFSI report.

The non-committal stance is then made explicit. The quotation above continues:

It is not the role of the *Publication Manual* to resolve these issues. 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. (APA, 2001, pp. 21-22)

It is unclear whether or how such equivocation is intended to modify clear and strong statements made elsewhere, for example those concerning the reporting of effect size measures and confidence intervals. The equivocation is all the more noteworthy, given that the Manual does make strong statements about such issues as margins and footnote styles.

Further confusion is added as the above quotation continues:

This edition attempts only to reflect the current views on the best practices with regard to data analytic approaches, reporting, and display. (APA, 2001, pp. 22)

The implication is that hypothesis testing—which appears frequently in the new Manual—is currently considered one of the best practices in psychology. Indeed there may be some situations in which hypothesis tests are best available practice given our current state of knowledge, however reformers have long argued that in a great many psychological applications NHST is one of the worst practices.

The section on statistical significance (APA, 2001, pp. 24-25) has been revised somewhat. It allows use of an a priori alpha level (“significance level” is introduced as a synonym), or reporting of a posteriori exact p values (referred to as “significance probabilities”), although the latter is preferred. It is also acceptable to use asterisks to indicate standard significance levels.

The TFSI’s crucial statement that “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” (TFSI report, p. 599) is ignored by the revised Manual. The Manual’s strong statement quoted above supporting confidence interval use appears alongside many paragraphs explaining NHST and how it should be presented. The new Manual includes many examples of NHST use, but not a single example of confidence interval use or of a recommended style for the presentation of confidence interval values. The unwritten message

continues to be “NHST business as usual”. There is no attempt to explain or justify the continued emphasis on NHST in examples nor is there any attempt to explain why the more detailed TFSI proposals are not included in the Manual.

From the perspective of statistical reform the new Manual must be judged extremely disappointing: It is a major opportunity missed. The good work of the TFSI has not had the influence in changing the APA’s official Manual that we would reasonably have expected, especially given its specific charter in relation to revision of the Manual.

Responses to reform proposals since 1994

Recent responses to both the TFSI report and the older, broader reform movement suggest that some journal editors have taken up the reporting of effect size measures as one important take-home message. In his guidelines to authors as editor-designate of Educational and Psychological Measurement, Thompson (1994) highlighted the problems of significance testing and mandated the reporting and interpretation of effect size measures. Other editors of APA and non-APA journals have mandated (Ellis, 2000; Heldref Foundation, 1997; Hresko, 2000; McLean & Kaufman, 2000; Murphy, 1997) or recommended (Kendall, 1997; Levant, 1992; Levin, 1995; Neeley, 1995) reporting of effect size measures.

Two problems with recent responses. We endorse the reporting of effect size, but have two concerns. First, effect size has been advocated largely to the exclusion of other aspects of the reform agenda. For example, without confidence intervals it is hard to estimate how accurate the estimated effect size is and this has serious implications for making substantive conclusions. Yet Ellis (2000) is the only editorial of those mentioned above that mentions confidence intervals for effect sizes. Statistical power, for example, although emphasized in the 1994 APA Manual, has not been mandated in these journals. The broader change in philosophy of analysis suggested by the TFSI is not reflected in recent editorial comments. Further, some editorial recommendations mandate reporting effect size measures when a NHST is reported. A simple-minded response to such recommendations would involve reporting effect sizes but still relying substantially on NHST.

Second, effect sizes are advocated as essentially unproblematic. We briefly mention two theoretical and practical problems. The use of standardized effect size measures has been questioned because of the effect of different measurement errors in different studies (e.g. Greenland, 1987). Although meta-analysts have discussed this problem of comparability across studies, it is not recognized widely in psychology’s reform literature. The second potential problem is poor interpretation of effect sizes. As we summarized elsewhere, empirical studies also show that where effect sizes are reported, they are rarely substantively

interpreted (Finch, Cumming, & Thomason, 2001; see also Keselman et al, 1998; Thompson, 1999; Vacha-Haase et al., 2000). Some editors aware of this problem (e.g. Thompson, 1994) have mandated interpretation of effect size measures. Analogous problems of interpretation arise for confidence intervals (Finch, Cumming, & Thomason, 2001).

In sum, effect sizes are selectively advocated without full discussion of potential pitfalls. Vacha-Haase et al. (2000) discussed this issue in relation to Cohen's suggested guideline values of small, medium and large effects: "Of course, if across disparate types of inquiries we used criteria for defining low, medium and large effects with the same rigidity that we have traditionally applied the .05 alpha level, we would merely be being mindless in another metric" (p. 422). This is the kind of limitation we identified historically in APA Manuals. Some current editorial recommendations represent little more than a new routine of calculating an effect size along with a NHST; reform must be much more substantial than the replacement of a flawed ritual (NHST) with any other narrow ritual.

The APA Manual

We, like others (e.g. Kirk, 2001) believe that reform requires advocacy and support from many sources. The APA Manual has a vital role to play. Budge and Katz (1995) described the Manual as "the primary resource on writing in the profession. It is also the single text which virtually every psychologist ... has contact with at some point in their career" (p. 218). Bazerman (1988) argued that "Today the American Psychological Association Publication Manual symbolizes and instrumentally realizes the influence and power of official style" (p. 259). He linked the development of official APA style and the Manual with the emergence of behaviourism in experimental psychology:

... the development of experimental psychology gives a particular interpretation to the experimental report that achieves a highly codified, institutionalized form. This codification stabilizes particular intellectual beliefs, empirical practices, and social relations around assumptions of a particular kind of research program (p. 259).

Budge and Katz (1995) also argued that the Manual guides "the reader/author ... to socialize them to the values of science done and presented in a certain way" (p. 233).

Psychologists have also commented on the influential role of the Manual. Kirk (2001), for example, emphasized:

The APA publication manual and similar manuals are the ultimate change agents. 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. (p. 217)

We note that other disciplines have no difficulty making simple statements requiring routine use of practices that psychologists have so far found difficult to adopt. For example, the Uniform Requirements for Biomedical Journals (International Committee of Medical Journal Editors, 1997) say “When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information” (p. 929).

The Manual's authoritative status does not mean that it is followed by all journals, editors or authors (Bazerman, 1988; Budge & Katz, 1995; Finch, Cumming, & Thomason, 2001). We have described how particular recommendations about statistical reporting practices are often not followed. We argue however that the Manual has shaped practice in a broad sense—by focussing on NHST to the exclusion of many other approaches. The Manual is, and will continue to be, very influential; whatever statements it makes about statistical methods and procedures will guide and shape future practice. For example, the APA in 1974 acknowledged the role of the Manual in shaping the content of psychology courses (APA, 1974, p. 5).

Development of the APA Manual guidelines on statistical methods is essential. It should be undertaken acknowledging the potential importance of whatever is said, along with what is not said. Any revision must have broader coverage of statistical methods, and should reflect the broader debate within psychology about NHST. Material describing the common misunderstandings of statistical matters, how these misunderstandings can (and have) damaged psychological research, and, in brief, the history of statistics in psychology should be included. As we argued above, the TFSI's report presents an important step in the right direction, but insofar as the fifth edition of the Manual (APA, 2001) did not carry the TFSI's proposals through, it is largely an opportunity missed.

Achieving reform

Full implementation of the fundamental changes recommended by the TFSI would constitute a substantial change in psychology's research culture. Securing such a change is vital for the future of the discipline; it requires conceptual and behavioral change by researchers across psychology's many sub-fields. We now briefly consider statistics education, and research that is needed to guide reform. These two issues are crucial for reform but have received insufficient attention by reform advocates.

Education for a change in research culture. NHST and its misconceptions permeate psychology so deeply that a considerable educational effort is required if present and future psychologists—researchers and practitioners—are to understand and adopt better ways to think about experimental design and research results. Major revisions to statistical education at undergraduate and graduate levels is needed (Thomason, Cumming, & Zangari, 1994), as well as support for current researchers to rethink what they have been doing. New textbooks and teaching software are required. Given that past statistics education in psychology has been problematic and that future psychologists will need a wider array of statistical skills, increased research on statistics learning is needed.

Research to guide reform. The reformers' claims that proposed new ways to analyze data and present results will be more easily and accurately understood are generally plausible, but in many cases there is little evidence to support them. For example, do researchers in fact have an easy intuitive grasp of what a confidence interval is showing? What is the best way to present a confidence interval? How can confidence intervals be used effectively in complex designs? How can their great potential to help researchers combine evidence across studies (Schmidt, 1996) be realized?

In our analysis of Journal of Applied Psychology reporting practices (Finch, Cumming, & Thomason, 2001) we found that even in the very few instances when confidence intervals are used most authors do not seem to understand their potential. Also, we have initial evidence that confidence intervals can be difficult to interpret (Fidler, Finch, Cumming & Thomason, 2001).

The TFSI report advocated the use, before conducting an experiment, of power estimates to guide experimental design including selection of sample size. In our view confidence intervals may often be more easily understood here, as well as later for the interpretation of results (Cumming & Finch, 2001). Cohen, for more than two decades the strongest advocate of the use by psychologists of power (Cohen 1962, 1969), later stated that confidence intervals can often with advantage be used in place of power (personal

communication to N.R. Thomason, 7 November 1994). An advantage of this suggested broadening of the role of confidence intervals is that avoiding the use of power would help reduce the role of NHST: Statistical power is inextricably enmeshed with NHST. Again, the proposal is little more than attractive speculation until evaluated empirically.

Most reformers are staunch experimentalists, and so it is all the more remarkable that they attribute so many psychological and pedagogic virtues to the confidence interval, with so little empirical support. It is clear that a comprehensive research program is needed. It should examine many things about confidence intervals, including how they can most effectively be pictured and reported, the role they should take in statistics education, the best relation with NHST, their role as estimates of precision to replace power, and the making of substantive interpretation of research findings. The research program should draw on experience in other disciplines, including medicine, where confidence intervals have been widely used for some time. More generally, any reform advocacy needs to be supported by empirical investigation of what is being advocated.

Editors, the gatekeepers

Some journal editors are responding to the call for reform, and some editors will publish papers that adopt alternative techniques. However, as we have showed elsewhere (Finch, Cumming, & Thomason, 2001) editors have implicitly endorsed APA recommendations selectively. Perhaps, given the mixed messages in the Manual (APA, 1994; 2001), this is unsurprising. A primary purpose of the Manual should be to give better and more comprehensive assistance to editors in their task of setting research standards.

Shrout (1997) argued that an editor instituting a “virtual ban” (p. 1) on NHST has in one medical case brought about substantial reform of reporting practices. The broad shift in philosophy of analysis proposed by the TFSI will not be achieved without strong editorial resolve, but more will be required. Manuscript reviewers, researchers, policy makers and statistics educators all have a responsibility to contribute, and the Manual must inform each of these constituencies.

Typically, journal editorials have not dealt extensively with statistical issues, but editors now have a crucial responsibility to support reform by providing clear and detailed policy statements about statistical methods for their area. They should give guidance on making best use of the Manual's requirements and advice, and support discussion of implementation issues as they arise. They can reward authors who provide striking examples of minimally sufficient analysis, and so give the psychological community good examples of reform practices.

Most broadly it is incumbent on every journal editor to find an effective combination of exhortation, education, support and sanction to ensure that published papers reflect good reformed statistical practice.

Conclusion

APA guidelines have in the past been little influenced by reformers' concerns and, although highly influential in a broad sense, have in many cases not been notably successful in making particular changes to researchers' statistical reporting practices. The latest Manual (APA, 2001) is, from a reform perspective, only a little better. Its continued emphasis on NHST in its examples is hard to understand, given some of its other statements.

Materials on quantitative methods in future APA Publication Manuals should provide psychologists with a richer and broader view of the application of statistics in their discipline. Extensive work is needed on new materials and curricula for statistical education. Research is needed on aspects of the new requirements, and critical analysis must be ongoing. Energetic work by editors and their gatekeeping associates is as vital as ever, and the APA must ensure that its editors succeed in their crucial leadership role. Such a broad range of approaches is required if psychology is to achieve a long overdue transformation—a culture change—in its statistical practice.

Beyond these important issues there remains the subsidiary but fascinating question wide open for study: Why, when other disciplines could years ago move beyond fixation on dichotomous NHST decision making, has psychology persisted with inefficient, damaging practices of statistical inference for so long?

References

Abelson, R. (1997). On the surprising longevity of flogged horses: why there is a case for the significance test. Psychological Science, 8, 12-15.

American Psychological Association. (1967). Publication Manual of the American Psychological Association (Rev. ed.). Washington, DC: Author.

American Psychological Association. (1974). Publication Manual of the American Psychological Association (2nd ed.). Washington, DC: Author.

American Psychological Association. (1983). Publication Manual of the American Psychological Association (3rd ed.). Washington, DC: Author.

American Psychological Association. (1994). Publication Manual of the American Psychological Association (4th ed.). Washington, DC: Author.

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

American Psychological Association, Council of Editors. (1952). Publication Manual of the American Psychological Association. Psychological Bulletin, 49(Suppl., Pt. 2), 389-450.

American Psychological Association, Council of Editors. (1957). Publication Manual of the American Psychological Association (Rev. ed.). Washington, DC: Author.

Anderson, J. E., & Valentine, W. L. (1944). The preparation of articles for publication in journals of The American Psychological Association. Psychological Bulletin, 41, 345-376.

Bakan, D. (1966) The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Bazerman, C. (1988). Shaping written knowledge: The genre and activity of the experimental article in science. Madison, Wisconsin: University of Wisconsin.

Beauchamp, K., & May, R. (1964) Replication report: Interpretation of levels of significance by psychological researchers. Psychological Reports, 14, 272.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, 70, 107-115.

Boring, E. G. (1919). Mathematical versus statistical significance. Psychological Bulletin, 16, 335-338.

Budge, G., & Katz, B. (1995). Constructing psychological knowledge: Reflections on science, scientists and epistemology in the APA Publication Manual. Theory & Psychology, 5, 217-231.

Carter, D. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

Carver, R. (1978). The case against significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Chandler, R. E. (1957). The statistical concepts of confidence and significance. Psychological Bulletin, 54(5), 429-430.

Chow, S. (1988). Significance test or effect size? Psychological Bulletin, 103, 105-110.

- Chow, S. (1996). Statistical significance: Rationale, validity and utility. London: Sage.
- Chow, S. (1998). Precis of Statistical significance: Rationale, validity and utility. Behavioral and Brain Sciences, 21, 169-239.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal & Social Psychology, 65, 145-153.
- Cohen, J. (1965). Some statistical issues in psychological research. In B. Wolman (Ed.), Handbook of clinical psychology. New York: McGraw-Hill.
- Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.
- Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and Psychological Measurement, 33, 107-112.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.
- Cohen, J. (1994). The earth is round ($p < 0.05$). American Psychologist, 49, 997-1003.
- Cortina, J. M., Dunlap, W. P. (1997). On the logic and purpose of significance testing. Psychological Methods, 2, 161-172.
- Craig, J., Eison, C., & Metze, L. (1976). Significance tests and their interpretation: An examination utilizing published research and omega squared. Bulletin of the Psychonomic Society, 7, 280-282.
- Cronbach, L. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. Educational and Psychological Measurement, 61, 530-572.
- Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. Psychological Bulletin, 63, 400-402.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70, 193-242.
- Ellis, N. (2000). Editorial. Language Learning, 50.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, *5*, 75-98.

Fidler, F., Finch, S., Thomason, N., & Cumming, G. (2001). Understandings and misunderstandings of confidence intervals: Confidence intervals as descriptive statistics. Manuscript in preparation.

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

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2001). Reform of statistical inference in psychology: The case of Memory & Cognition. Manuscript in preparation.

Fisher, R. (1935). The design of experiments. Edinburgh, Scotland: Oliver & Boyd.

Frick, R. (1995). Accepting the null hypothesis. Memory & Cognition, *23*, 132-138.

Frick, R. (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.

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. Hillsdale, NJ: Erlbaum.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. Educational Researcher, *5*(10), 3-10.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage.

Grant, D. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, *60*, 54-61.

Grayson, D., Pattison, P., & Robins, G. (1997). Evidence, Inference and the “Rejection” of the Significance Test. Australian Journal of Psychology, *49*(2), 64-70.

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. Epidemiology Review, *9*, 1-30.

Hagen, R. (1997). In praise of the null hypothesis statistical test. American Psychologist, *52*, 15-24.

Hammond, G. (1996). The objections to the null hypothesis as a means of analysing psychological data. Australian Journal of Psychology, *48*, 104-106.

Harlow, L. (1997). Significance testing introduction and overview. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), What if there were no significance tests? Mahwah, NJ: Erlbaum.

Harlow, L., Mulaik, S., & Steiger, J. (1997). What if there were no significance tests? Mahwah, NJ: Erlbaum.

Harris, R. (1997). Significance tests have their place. Psychological Science, 8, 8-11.

Hays, W. (1963). Statistics for Psychologists. New York: Holt, Rinehart & Winston.

Heldref Foundation (1997). Guidelines for contributors. Journal of Experimental Education, 65, 95-96.

Hotelling, H., Bartky, W., Deming, W., Friedman, M., Hoel, P. (1948). The teaching of statistics. Annals of Mathematical Statistics, 19, 95-115.

Hresko, W. (2000). Editorial policy. Journal of Learning Disabilities. 33, 214-215.

Hunter, J. E., & Schmidt, F. (1990). Methods of meta-analysis: correcting error and bias in research findings. Newbury Park: Sage.

Hunter, J. E., Schmidt, F., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Instructions in regard to the preparation of manuscripts (1929). The Psychological Bulletin, 26, 57-63.

International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. Journal of the American Medical Association, 277, 927-934.

Kendall, P. C. (1997). Editorial. Journal of Consulting and Clinical Psychology, 65, 3-5.

Kern, G. & Lewis, C. (1979). Partial omega squared for ANOVA designs. Educational and Psychological Measurement, 39, 119-128.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350-386.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61, 213-218.

LaForge, R. (1967). Confidence intervals or tests of significance in scientific research. Psychological Bulletin, 64, 446-447.

Levant, R. F. (1992). Editorial. Journal of Family Psychology, 6, 3-9.

Levin, J. R. (1995). Editorial.: Journal alert! Journal of Educational Psychology, 87, 3-4.

Loftus, G. R. (1993a). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. Behavior Research Methods, Instruments & Computers, 25(2), 250-256.

Loftus, G. R. (1993b). Editorial comment. Memory & Cognition, 21(1), 1-3.

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(6), 161-171.

Lykken, D. (1968). Statistical significance in psychological research. Psychological Bulletin, 70, 151-159.

Maxwell, S., Camp, C., & Avery, R. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.

McLean, J., & Kaufman, A. (2000). Editorial: Statistical significance testing and Research in the Schools. Research in the Schools, 7(2).

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting & Clinical Psychology, 4, 806-843.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. Psychological Reports, Monograph Supplement 1-V66.

Morrison D. E., & Henkel R. E. (1970). The significance test controversy - a reader. London: Butterworths.

Mulaik, S., Raju, N., & Harshman, R. (1997). There is a time and place for significance testing. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), What if there were no significance tests? Mahwah, NJ: Erlbaum.

Murphy, K. R. (1997). Editorial. Journal of Applied Psychology, 82, 3-5.

Neeley, J. H. (1995). Editorial. Journal of Experimental Psychology: Learning, Memory & Cognition, 21, 261.

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641-650.

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioural sciences. Chichester: Wiley.

- O'Grady, K. (1982). Measures of explained variance: cautions and limitations. Psychological Bulletin, *92*, 766-777.
- Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. The Journal of Psychology, *55*, 33-38.
- Rosenthal, R. & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. Psychological Reports, *15*, 570.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, *44*, 1276-1284.
- Rozeboom, W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, *57*, 416-428.
- 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. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, *1*, 115-129.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the statistical power of studies? Psychological Bulletin, *105*, 309-316.
- Shrout, P. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. Psychological Science, *8*, 1-2.
- Thomason, N., Cumming, G., & Zangari, M. (1994) Understanding central concepts of statistics and experimental design in the social sciences. In K. Beattie, C. McNaught & S. Wills (Eds.), Interactive multimedia in university education: Designing for change in teaching & learning. Amsterdam: Elsevier.
- Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, *54*, 837-847.
- Thompson, B. (1996). AERA Editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, *25*(2), 26-30.
- Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. Educational Psychology Review, *11*, 157-169.
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work. American Psychologist, *24*, 83-91.
- Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, *76*, 105-110.

Vacha-Haase, T., & Ness, C. N. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and Practice, Professional Psychology: Research and Practice, 30(1), 104-105.

Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and uses in MECD. Measurement and Evaluation in Counseling and Development, 31, 46-57.

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.

Wilkinson, L., & Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: guidelines and explanations. American Psychologist, 54, 594-604.

Wilson, W., & Miller, H. (1964). A note on the inconclusiveness of accepting the null hypothesis. Psychological Review, 71, 238-242.

Wilson, W., Miller, H., & Lower, J. (1967). Much ado about the null hypothesis. Psychological Bulletin, 64, 188-196.

Author notes

The authors acknowledge the support of the Australian Research Council. Bruce Thompson provided helpful comments on a draft.

Correspondence about this article should be addressed to Geoff Cumming, School of Psychological Science, La Trobe University, Bundoora, Australia, 3086. Email: g.cumming@latrobe.edu.au

Keywords

statistical significance, APA Manual, reform of statistical inference, publication guidelines, confidence intervals

Biographical notes

Sue Finch works at the Heymans Institute for Psychological Research at The University of Groningen. She has a PhD in psychology. Her research interests include finding better

ways of teaching and learning statistics, use of multimedia for statistics education, and reform of statistical practices in the social sciences.

Neil Thomason has a doctorate in the philosophy of science under Paul Feyerabend. He has long been interested in the socio-psychology of scientific rationality -- how not-always-terribly-rational individuals en masse regularly produce rational understanding of a subject. The history of Null Hypothesis Testing in psychology provides a fascinating case where this did not happen.

After a first degree in mathematical statistics at Monash University, Melbourne, Geoff Cumming took his DPhil in experimental psychology at Oxford University. He has been in the School of Psychological Science at La Trobe University, Melbourne, for many years. His recent research interests include interactive multimedia for statistics education, and reform of research practices.