

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.

© Sage Publications

Journal website: <http://www.sagepub.com/journal.aspx?pid=165>

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

Reporting of statistical inference in the Journal of Applied Psychology: Little
evidence of reform

Sue Finch¹, Geoff Cumming¹ & Neil Thomason²

1 = La Trobe University, Melbourne, Australia

2 = The University of Melbourne, Melbourne, Australia

Abstract

Reformers have long argued that misuse of Null Hypothesis Significance Testing (NHST) is widespread and damaging. We analyzed 150 papers from the Journal of Applied Psychology (JAP) covering 1940 to 1999. We examined statistical reporting practices related to misconceptions about NHST, APA guidelines, and reform recommendations. Our analysis reveals (a) inconsistency in reporting alpha and p-values, (b) use of ambiguous language in describing NHST, (c) frequent acceptance of null hypotheses without consideration of power, (d) that power estimates are rarely reported, (e) virtually no confidence intervals. APA guidelines have been followed only selectively. Research methodology reported in JAP has increased greatly in sophistication over 60 years, but inference practices have shown remarkable stability. There is little sign that decades of cogent critiques by reformers had by 1999 led to changes in statistical reporting practices in JAP.

Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform

The use of Null Hypothesis Significance Tests (NHST) in psychology has persisted despite a long history of trenchant criticism of the logic of the procedure and of psychologists' use and understanding of it (e.g. Bakan, 1967; Berkson, 1938; Cohen, 1962, 1990, 1994; Grayson, Pattison & Robins, 1997; Hammond, 1996; Harlow, Mulaik & Steiger, 1997; Hunter, 1997; John, 1992; Lykken, 1968; Schmidt, 1992, 1996; Tversky & Kahneman, 1971). Reformers have argued that NHST is detrimental to theory development and has damaged psychological research (e.g. Oakes, 1986; Meehl, 1978; Rossi, 1997). Among notable reform recommendations are the reporting of effect sizes and confidence intervals, reduced reliance on NHST, reporting of statistical power, routine use of exploratory data analysis, and an emphasis on the substantive interpretation of findings.

We agree that reform is important and long overdue. To assess the progress of reform over the last 60 years we document reporting practices in a leading American Psychological Association (APA) journal--the Journal of Applied Psychology (JAP). We examined how statistical inference has been reported, with particular attention to practices consistent with APA guidelines, and those consistent with reform recommendations. Our analysis of 150 empirical papers from JAP spans from the early rapid increase in the use of NHST (1940 & 1955) through several decades of increasing call for change (1970 to 1999). As a small test for generality we also analyzed thirty 1999 papers from the British Journal of Psychology (BJP).

Below we briefly review the uptake of NHST in psychology, and some of the literature on misconceptions associated with NHST, before outlining the changing

APA standards for reporting inference. We then review other studies of reporting practices in psychological research journals.

The uptake of NHST in psychology

Gigerenzer and Murray (1987) described the approach to data before 1940: “The overall picture is an extensive, piecemeal, and above all nonstandardized presentation of descriptive statistics, and a comparatively flexible and negotiable attitude toward the issue of inference from data to hypothesis” (p.19). Two inferential indexes, the probable error and the critical ratio, were often reported although they were used in different ways by different authors.

Gigerenzer and Murray (1987) described an inference revolution in psychology, taking place between 1940 and 1955. Statistical techniques for testing hypotheses that had been developed by Ronald A. Fisher, and Jerzy Neyman and Egon Pearson, were enthusiastically embraced. By 1955, a confused hybrid version of Fisher’s and Neyman-Pearson’s approaches to inference had “revolutionized American research practice” (Gigerenzer & Murray, p. 22). Uptake of this hybrid had many consequences for psychology, not least of which were persistent confusions and misinterpretations of the results of NHST by students, textbooks and researchers.

Misconceptions about NHST: alpha, p-value, and significance level

The p -value is the calculated probability of the observed result—or results that deviate more from the null hypothesis—for the given sample size, assuming the null hypothesis is exactly true in the population. The ‘significance level’ is the criterion set for rejecting the null hypothesis; in Neyman-Pearson terms it is also ‘alpha’ or the Type I error rate (the a priori probability of rejecting the null hypothesis when it is, in fact, true).

Misconceptions about these concepts are widespread. Low p -values are incorrectly interpreted as quantifying the improbability of the null hypothesis (Falk & Greenbaum, 1995). Also, the p -value is taken to be an inverse indicator of effect size and importance of a finding (Thompson & Snyder, 1998), and of replicability (Gigerenzer, 1993; Oakes, 1986).

Problems with the term ‘significance level’ may arise partly from differences in usage. For example, Neyman and Pearson used the term to refer to the long-term relative frequency of Type I errors, while Fisher used it at one time to refer to a standard for rejecting a null hypothesis and at other times to refer to the p -value. We will avoid further use of the term ‘significance level’.

Type I error terminology does not make sufficiently clear the conditional nature of the probability. The Type I error probability is often incorrectly taken to be the probability of a Type I error conditional on rejection of the null hypothesis, rather than on the null being true (Pollard, 1993; Pollard & Richardson, 1987).

Mindful of these many widespread misconceptions, in our analysis of journal papers we examined how researchers report p -values and alpha levels. We identified use of exact and relative p -values, reporting of p -values relative to more than one standard, and a priori specification of the alpha level. This painted one part of the picture of how researchers use NHST.

Misconceptions about NHST: interpretation of results

Reformers have identified the widespread use of a .05 alpha level as a dichotomous decision criterion as a central problem (Loftus, 1996; Rosnow & Rosenthal, 1989). Statistically significant (ss) results are often interpreted as having substantive meaning, without consideration of effect size or practical meaning of the finding. In contrast, a statistically non-significant (sns) result can “mean ruin, despair,

and ... suddenly thinking of a new control condition that should be run” (Rosnow & Rosenthal, 1989, p.1127). Such a response ignores the plausible possibility that there is a real effect and the sns result simply reflects low statistical power.

Meta-analysts have shown most dramatically the damaging effects of relying on a dichotomous decision criterion (e.g. Hunter & Schmidt, 1990). Narrative literature reviews based on counts of ss and sns findings can be highly misleading. As Hunter and Schmidt (1990) concluded: “The typical use of significance test results leads to terrible errors in review studies. Most review studies falsely conclude that further research is needed to resolve the ‘conflicting results’ in the literature” (p.31).

Defensible interpretation of sns results must consider sample size, the extent of sampling variability, and power: ss is often not achieved simply because of low power. Conversely, a sns result in a very high power experiment may constitute reasonable grounds for concluding that the null hypothesis is true. Consonant with the meta-analysts’ observation that psychologists frequently disregard these vital issues, Tversky and Kahneman (1971) presented evidence of psychologists’ tendency to choose a substantive explanation for the difference between two results when a larger study provided a ss result and a smaller study did not. Their persuasive conclusion was that even research psychologists are insensitive to the relationship between sample size and sampling variability (Tversky & Kahneman, 1971).

One reason for problems of NHST interpretation may be differences between Fisher and Neyman-Pearson views about sns results. Fisher initially allowed that the null hypothesis could be disproved but not proved; sns results were of little interest (Gigerenzer, 1993). Later he allowed that sns results might have some bearing on the null hypothesis, but gave little indication of what specific interpretation is justified. In

the Neyman-Pearson framework two precise competing hypotheses are specified and so power can be calculated.

It is clear that psychologists, like many others, have difficulty in appropriately using and interpreting NHST. In the current study we examined aspects of authors' journal reporting practices that relate to NHST issues mentioned above, especially interpretation of sns results and recognition of the role of sample size.

Reformers have written cogently about the problems of NHST, but they cannot mandate the reporting practices psychologists should adopt. We next consider the official standards for journal reporting practice by first examining standards set by the APA, and then reviewing JAP editorial policies.

APA standards for reporting

In Finch, Thomason and Cumming (2000) we presented a discussion of APA reporting standards over the years. Here we will first summarize the recommendations for reporting statistical information made in the Publication Manual of the APA (APA, 1952, 1957, 1967, 1974, 1983, 1994), then consider the report of the APA Task Force on Statistical Inference (TFSI), which undertook to advise on future APA guidelines (Wilkinson & TFSI, 1999, henceforth TFSI report).

The first detailed instructions for reporting statistical inference appeared in the second edition of the Manual (APA, 1974). That Manual stated: "The results should summarize the collected data and your statistical treatment of them." (p.18). Reports of statistical tests were to include the test statistic, degrees of freedom, and p-value. The examples showed that the relative p-value was intended: the p-value relative to some standard for statistical significance (e.g. $p < .05$). The "statistical significance levels, if any" (p.15) should be stated in the abstract (also APA, 1967). There was no requirement to include measures of central tendency or spread. In the third edition of

the Manual (APA, 1983) the instructions relating to statistical information were essentially the same as those in the 1974 edition, however the inclusion of descriptive statistics (means and standard deviations) was advised.

The fourth and current edition (APA, 1994) provided more extensive recommendations. It detailed some statistics that researchers should include when reporting different inferential test results. For the first time, a recommendation to specify the a priori alpha level was made. For the first time, the difference between the alpha level and the p-value was explained, and the reporting of exact p-values suggested. The statement that the abstract should include 'significance levels' was retained. Most importantly, for the first time authors were encouraged to report effect sizes and to "take seriously the statistical power" (p.16); the need to consider power in interpreting sns results was discussed.

Recently, the TFSI report made recommendations that covered many aspects of research design, analysis and interpretation, and that recognized many reform issues. In the current study we examined several aspects of reporting practice discussed in the TFSI report, including use of confidence intervals, consideration of sample size and statistical power, interpretation of NHST results, and acceptance of a null hypothesis.

Official JAP standards for reporting

Editorial commentary in the JAP since the early 1980s has indicated some official concern about a variety of statistical issues. In a lengthy review of his period as editor from 1976 to 1982, Campbell (1982) discussed reasons for manuscript rejection and wrote that low power was one of the less frequent disqualifiers, although he did not report how often power calculations were available. He further commented:

... it is true that there is an evaluation asymmetry between significant and nonsignificant results. Besides the lack of any true relationship among the latent variables, the decision not to reject the null hypothesis can be a function of lack of power, lack of validity for the measures, unreliable measurement, lack of experiment control, and so on. Studies that wish to give a substantive interpretation to negative results must be reasonably well done. (p.693)

Campbell also devoted a paragraph to the problems of NHST, emphasizing that p-values are not a measure of substantive significance:

Perhaps p values are like mosquitoes. They have an evolutionary niche somewhere and no amount of scratching, swatting, or spraying will dislodge them. Whereas it may be necessary to discount a sampling error explanation for results of a study, investigators must learn to argue for the significance of their results without reference to inferential statistics. (p.698)

The next editor, Guion (1983), noted the importance of effect size measures:

How substantial are the reported effects? Too often we receive manuscripts reporting significance levels but saying nothing at all about the sizes of the effects. An estimate of effect size not only helps readers and reviewers evaluate the contribution an article makes, but it may be indispensable for future meta-analyses. (p.548)

He was also concerned about a tendency for the selective reporting of just those results that reach statistical significance.

Schmitt (1989) advocated the use of simple methods of data analysis, where appropriate, to support better communication of results. Bobko (1995) reiterated this view, as did Murphy (1997).

Murphy (1997) also strongly encouraged reporting of effect sizes with significance test results:

The Publication Manual ... 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 specific justification for why effect sizes are not reported. (p.4)

While commendable, this passage illustrates a tendency of JAP editors to endorse Publication Manual recommendations selectively. Murphy does not mention that the Manual also recommended, for example, that researchers report a priori alpha, and take seriously statistical power.

In summary, various JAP editors have discussed problems of reliance on NHST (Campbell, 1982; Guion, 1983), misinterpretation of statistically non-significant results (Campbell, 1982), the need to report effect sizes (Guion, 1983, Murphy, 1997), the importance of using simple analytic techniques (Bobko, 1995; Murphy, 1997; Schmitt, 1989) and the need to provide information for future meta-analyses (Guion, 1983). Campbell's (1982) discussion of negative results and statistical power implies that low power was not a great concern to him. It is clear that at least two recent JAP editors appreciated at least some of the major reform issues, but have their concerns been reflected in the papers published under their editorial guidance?

In the following two sections we discuss briefly some studies of journal reporting practices of most relevance to our own analysis.

Previous studies of JAP reporting practices

Hubbard, Parsa and Luthy (1997) tracked the uptake of NHST in the JAP from 1917 to 1994. They documented an increasing reliance on significance tests, with

NHST appearing in over 90% of articles in the 1990s. Increasing and apparently enthusiastic adoption of NHST was noted during 1940-1955: NHST became the standard in the JAP, as in many other journals.

There are several indications that the statistical power of research reported in the JAP may on average be relatively high. As noted above, Campbell (1982) judged that low power was not a major problem.

Chase and Chase (1976) analyzed 121 articles published in JAP in 1974 and concluded that the “statistical power exhibited in this sample of applied psychological research was relatively high” (p.236), compared with power estimates from other sub-disciplines of psychology including Cohen’s (1962) study of the Journal of Abnormal and Social Psychology. They noted, however, that even in the JAP average power approached Cohen’s (1965) recommended level of .80 only for large effect sizes. Muchinsky (1979) sampled 867 JAP articles at 5 year intervals from 1957 to 1977 and found that average sample sizes for papers reporting field studies ranged between 353 and 610. Muchinsky claimed that such large sample sizes are likely to have high statistical power.

Kirk (1996) studied the use of measures of effect size in 1995 volumes of JAP and three other APA journals (Journal of Educational Psychology, Journal of Experimental Psychology, Learning & Memory, and the Journal of Personality & Social Psychology). Across the four journals, the percentage of articles that used inferential statistics and reported effect size measures ranged from 22% to JAP’s 77%. This suggests that JAP authors are most likely to follow this reform practice, but Kirk noted:

Before anyone concludes that authors of articles in the Journal of Applied Psychology are more aware of the limitations of [NHST], remember that these

authors are more likely to use regression and correlation procedures. Computer packages routinely provide [effect size measures] for these procedures. (p.752)

These studies found that some reform practices were reflected to some extent in JAP – largish sample sizes, reporting effect sizes and somewhat greater power than in some other psychology journals. A major aim of the current study is to extend the examination of JAP papers to a larger range of reform issues.

Studies of reporting practices in other psychological journals

Four recent studies of reporting practices in other journals are relevant: These are studies of the Journal of Counseling and Development (25 articles from 1995; Thompson & Snyder, 1998), the Journal of Experimental Education (22 articles from 1994-1995; Thompson & Snyder, 1997), Measurement and Evaluation in Counseling and Development (68 articles from 1990-1996; Vacha-Hasse & Nilsson, 1998), and Professional Psychology: Research and Practice (265 articles from 1990-1997; Vacha-Hasse & Ness, 1999).

Standardized effect sizes (e.g. r^2 , η^2 , ω^2 , Cohen's d) were reported in upwards of one quarter of the articles examined in these studies (Thompson & Snyder, 1997, 1998; Vacha-Hasse & Nilsson, 1998; Vacha-Hasse & Ness, 1999). The Journal of Experimental Education most frequently reported standardized effect sizes. As Thompson and Snyder (1997) noted, this followed articles in the journal in 1993 that emphasized reporting of effect sizes and so “one might expect disproportionately more reports in this journal” (p.79). However authors rarely interpreted the effect sizes they reported (Thompson & Snyder, 1997, 1998).

Vacha-Hasse and Nilsson (1998) examined reporting of NHST: Most papers (84%) followed the APA (1994) guidelines for reporting inferential tests in that they included degrees of freedom, the value of the test statistic and the (relative) p -value,

but only 13% reported a priori alpha levels. Vacha-Hasse and Ness (1999) came to a similar conclusion.

All four studies examined the language used to describe ss (and sns) results. The term ‘significant’ was often used where ‘statistically significant’ was meant (Thompson & Snyder, 1997, 1998). The term ‘statistical significance’ was used appropriately in only 13% (Vacha-Hasse & Ness, 1999) and 34% of articles (Vacha-Hasse & Nilsson, 1998). As Thompson and Snyder (1998) emphasized, “it can be confusing if ‘significant’ is used within the same article in some places to mean ‘important’ and in other places to mean ‘statistically significant’” (p. 439). This is an old issue: Boring (1919) emphasized the distinction between mathematical significance and substantive importance. Thompson (1994, 1996) recommended the use of the precise term ‘statistical significance’, as did Carver (1993). We took up this issue in our current study.

The analyses of reporting practice mentioned above indicate that in some sub-disciplines many authors reported effect size while in other sub-disciplines they did not. In any case, they usually did not use effect size to assist in interpretation. Further, the studies suggest that problems of interpretation and language are pervasive.

Sixty years of reporting practice in JAP

We selected JAP as a long-established, leading APA journal that publishes empirical research in a wide range of fields. Also, as noted above, some aspects of reporting practice in JAP have already been documented. As a small step towards assessing generality we examined articles from the 1999 volume of BJP. We chose the BJP as an international journal of broad scope whose format is similar to that of the JAP, although its guidelines to authors do not refer to the APA Manual.

We noted at various points above most of the practices we chose for analysis and the rationale for our choice. In addition we studied two important reform recommendations that seem not to have been much investigated: use of confidence intervals, and visual representations of measures of variability.

Method

Sample of papers

We chose from JAP the first 30 papers that included any report of data and statistical inference from each of the years 1940, 1955, 1970, 1985 and 1999. Papers reporting probable error or critical ratios in 1940 were included, following the criteria of Hubbard et al. (1997). Similarly, 30 papers published in BJP in 1999 were selected.

Choice of reporting practices

After considering reform issues, disciplinary recommendations for reporting, previous studies of JAP reporting practices and the literature on statistical misconceptions, we identified an initial list of reporting practices to study. We developed descriptions of these practices and identified examples from papers. We discussed and refined the descriptions, then carried out trials of independent coding. Although prototypical examples of many interpretive practices could be identified, for a number of practices it was difficult to achieve consensus on the criteria. Variability and ambiguity in the language used by authors to describe and interpret results was often a major source of difficulty. We therefore reduced the number of practices to be coded: We selected practices that were central to reform, and able to be reliably identified by coders.

We chose 12 practices, in accord with the rationale in the Introduction above, and aiming for a balance between the number of practices, the number of papers to be reviewed, and the need for high coding reliability.

Alpha and p-value practices

Practice 1: p-value as level of confidence. Referring to the p-value as the 'level of confidence'. This ambiguous language may support overinterpretation of the meaning of the p-value (Gigerenzer, 1993).

Practice 2: exact and relative p-values. Reporting of exact ($p = .03$) and relative p-values ($p < .05$). Reporting of exact p-values was first suggested in the 1994 APA Manual, and may indicate a move away from the use of .05 as a dichotomous decision criterion. We also noted when p-values are reported in the abstract of articles from 1970 onwards, as recommended in APA Manuals since 1967. (Abstracts were not routinely published in 1940 and 1955.)

Practice 3: a priori alpha. Explicit specification of an a priori alpha level. The 1994 APA Manual stated: "Before you begin to report specific results, you should routinely state the particular alpha level you selected" (p.17).

Practice 4: more than one standard for statistical significance. Use of more than one standard for ss. Reports of several relative p-values (e.g. $p < .05$ and in the same paper $p < .01$) imply that more than one alpha level is being used.

Statistically significant and non-significant results

Practice 5: 'statistical significance' terminology. Use of the unambiguous term 'statistically significant'; we recorded this separately for ss and sns findings. Our criteria included cases where: (a) the term 'statistically significant' was stated explicitly and (b) the term 'significant' was accompanied in the same sentence by information about a statistical test; this made clear the intended meaning of

‘significant’. An illustration of (a) would be: “The main effect of gender was statistically significant ...” An example of (b) would be: “The main effect of gender was significant ($F(1, 39) = 10.4, p < .05$).”

Practice 6: ambiguous 'significance' terminology. Use of the ambiguous term ‘significant’ (rather than the unambiguous term ‘statistically significant’). Again, we recorded ss and sns cases separately. If the term ‘significant’ is used, and the criteria for Practice 5 are not met, the term could be interpreted as implying either statistical significance or practical importance. An illustration would be: “The effect of gender was significant.”

Practice 7: non-report of test statistic. Failure to report the numerical value of a test statistic. This practice was recorded separately for ss and sns cases. An inferential statistic was considered missing if there was an unambiguous statement of statistical significance (Practice 5) but the test statistic was not stated in the text or elsewhere (e.g. footnote, table, graph). For example: “The main effect of gender was not statistically significant.”

Practice 8: discounting statistical non-significance. A sns result is discounted: A sns result is described or interpreted as if it had been a ss result. This includes cases when a sns ‘trend’ in the results is given a substantive interpretation. For example: “Although the main effect of gender was not statistically significant, the females clearly performed better than the males.” Cases where results are described as marginally significant or approaching significance, although p-values did not meet the standard for ss otherwise used in the paper, were also included. For example: “The main effect of gender was marginally significant ($p < .06$).”

Practice 9: accepting a null hypothesis. A sns result is interpreted as providing evidence that the null hypothesis is true. The criteria for this practice required a

reference to a statistical test that was directly linked with a statement that there were no differences or effects. For example: “The resulting value was non-significant, $t(41) = .83$ which indicates that males were no different from females in their response to stress.” Proper interpretation of sns results should consider statistical power or the width of the confidence interval.

Sample size, power, and confidence intervals

Practice 10: sample size and power. Awareness of the role of sample size in NHST. We recorded any indication at all that authors were aware of the importance of sample size in NHST or of the notion of power. This is a very broad definition and includes, for example, any comment about the effect of small sample sizes on NHST results. We also noted cases where statistical power was calculated (either a priori or post hoc).

Practice 11: confidence intervals. Reporting confidence intervals in any form. The routine reporting of confidence intervals has been almost universally recommended by reformers including the TFSI.

Practice 12: visual display of measures of variations. Visual display of measures of variation (standard errors, standard deviations, confidence intervals). This includes error bars on graphs. Reporting of visual representations of data variability is recommended by reformers and TFSI.

Coding procedure and reliability

We examined each paper and noted whether or not we could find a single example of a practice; further examples of the same practice were not recorded. All papers were coded by the first author. The second author independently cross-coded a 10% sample from each year. All differences between the two coders arose from occasional missed information, rather than differences in application of the criteria.

In total, 107 practices were identified as present in the 15 papers. The first author failed to identify 5 (less than 5%) of these practices; the missed practices were distributed across years and types of practice. (These are included in the data presented below.) We thus judged coding reliability to be very satisfactory, although the results reported below may be slight underestimates.

Results

The results are based on 30 JAP papers in each year, except that just 27 papers in 1940 and 29 papers in 1970 included at least one sns finding, and 29 papers in 1955 included at least one ss finding. Hence all but five of the 150 JAP papers included both ss and sns findings, no doubt reflecting the fact that papers usually report many tests. Papers with at least one ss finding are referred to as ‘ss papers’, and papers with at least one sns findings are called ‘sns papers’. Figures 1 to 7 show most of the results. Each figure shows the proportion of papers in a year with at least one occurrence of a practice; the standard error of the proportion is also shown. Doubling the standard error gives the half-width of an approximate 95% confidence interval for the proportion. Also shown are the results for the 30 BJP papers from 1999 (29 ss papers, 29 sns papers). In the Results and Discussion, we have included examples from JAP papers but, to preserve the anonymity of authors, we have omitted reference details: This information is available from the first author.

Alpha and p-values

Practice 1: Only authors in 1955 (33%) and 1970 (10%) described the p-value as a level of confidence.

Practice 2: Figure 1 shows that reporting of exact p-values in the JAP declined since 1940 and in recent decades has remained markedly low. By contrast, in 1999

two-thirds of BJP articles reported exact p -values. Figure 1 also shows that over 90% of papers since 1955 have reported relative p -values; in the 1985 and 1999 JAP samples the figure was 100%. The only reports of p -values in JAP abstracts were in 1985 (10%); there were none in the 1999 BJP abstracts.

Practices 3 and 4: Alpha was specified a priori in between 7% and 17% of papers—the highest level being in 1940 (Figure 2). The rate in 1999 was little changed from earlier rates, despite the recommendation of the 1994 APA Manual.

By contrast, implicit use of more than one alpha level in JAP averaged at 77% in the period from 1955 to 1999, after being low in 1940 (Figure 2). In BJP in 1999 the figure was almost 100%.

Statistically significant and non-significant results

Practices 5 and 6: Figure 3 shows for ss results the proportion of ss papers in which just one of the terms ‘statistically significant’ or the ambiguous ‘significant’ is used, and that in which both terms are used. Figure 4 gives the proportions for sns results. The proportion of JAP articles using both unambiguous and ambiguous terms appears to increase somewhat over time, both for ss (Figure 3) and sns results (Figure 4), and since 1955 the proportion is generally higher for ss than sns. Accurate use of language—use of just the unambiguous term—is infrequent in the JAP, being about 10% for each of ss and sns in 1999; it is higher in the BJP. Since 1955, use of the ambiguous term alone is higher for sns than ss results, especially in recent years.

Practice 7: Figure 5 shows that failure to report an inferential statistic when an unambiguous statement of ss or sns is made becomes less frequent over time. However over 20% of the JAP papers from 1999 that include one or more unambiguous statements of statistical significance (or non-significance) do not include a numerical test statistic.

Practices 8 and 9: Figure 6 shows that the proportion of JAP sns papers in which a sns result is discounted ranges between 10% for 1970 and 27% for 1999; for 1999 BJP papers the figure is 41%. Figure 6 also shows that the proportion of sns papers in which a null hypothesis is accepted averages 38% ; it is still 37% for JAP in 1999.

Sample size, power, and confidence intervals

Practice 10: In the 1940 sample, 30% of JAP papers make at least some mention of the importance of sample size or (more rarely) statistical power. As Figure 7 shows the percentage drops in 1955 and 1970, but rises to around 40% in 1985 and 1999. However, no JAP paper before 1985 provides information that suggests that power had been calculated and only 3 (10%) of JAP papers articles from 1985 and 3 from 1999 provide information that suggests that power had been calculated. For example, “Power analyses conducted on the statistical tests indicate that the main effects ... could have been detectable given an alpha = .05, and the [sample sizes] involved.” One of the three 1985 papers provides an estimate of power (a priori); all three 1999 papers give an estimate (2 a priori, 1 post hoc). In BJP from 1999, 9 (27%) papers mention power or the role of sample size; just one (3%) reports the power estimate. Published papers provide little evidence that authors take power seriously as recommended in the 1994 Manual.

Practices 11 and 12: Of the 150 JAP papers only 4 contain reports of confidence intervals (1955, 1985, 1985, 1999) and only two represent variation in data visually (1955, 1999). In BJP in 1999, only 1 paper (3%) reports confidence intervals, and 6 (20%) include a visual representation of variability.

Discussion

Little improvement in many years

Psychology's research and analysis techniques have advanced dramatically: In 1940, studies often relied on simple comparisons of means or frequencies, but by 1999 many sophisticated techniques were being used, including complex multivariate analysis, path analysis and meta-analysis. It is therefore all the more remarkable that our analysis shows many aspects of reporting practices have remained surprisingly stable during the half century since the inference revolution. A disappointing corollary is that we found little sign of reform practices, even in 1999.

We will first look more closely at each reporting practice and then consider how use of confidence intervals might help avoid the problems we identified in the ways statistical inference is reported. Finally we examine the implications of our results for the reform process.

Alpha and p-values

Reporting practice has been remarkably consistent since 1955 (Figures 1 & 2). The great majority of authors report p-values relative to two or more standards. A priori alpha levels are rarely provided and since 1940 very few JAP authors report any exact p-values. In earlier years authors may have had difficulty calculating exact p-values, but these are now provided routinely by analysis software. BJP (1999) authors tend to include both relative and exact p-values; we do not know why BJP papers include exact p-values so much more often than JAP papers. We noted, although we did not code for this, that p-values are often not reported at all for sns results, rather the shorthand 'ns' is used.

The failure to specify alpha a priori may arise from widespread acceptance of .05 as the standard level and is further evidence of routine use of NHST. Consistent

with this, $p < .05$ is the most frequently reported relative p -value. Typical practice is to report relative p -values using two or more implicit alpha levels, most commonly .05 and .01, sometimes .001. This use of more than one implicit alpha level might be seen as a positive move away from a dichotomous decision criterion, but the practice is potentially problematic.

Use of several relative standards is taken by some psychologists to imply different standards of research quality (Gigerenzer, 1993; Loftus, 1993): one star ($p < .05$), two star ($p < .01$) or even three star ($p < .001$). Gigerenzer, for example, described how Melton considered the p -value in making decisions for accepting manuscripts in the Journal of Experimental Psychology. Articles with $p > .05$ were unlikely to be published. Even if $p < .05$ the article might not be accepted, but if $p < .01$ publication was more likely. Relying heavily on the p -value as an indicator of research quality diverts attention from important issues including effect size, amount of variability in the dependent variables, and sample size—let alone the appropriateness and precision of the design and methodology.

Reporting of exact p -values should help reduce the reliance on statistical tests as dichotomous (or trichotomous) decision criteria. As Rosnow and Rosenthal (1989) wrote: "... there is no sharp line between a 'significant' and a 'nonsignificant' difference; significance in statistics, like the significance of a value in the universe of values, varies continuously between extremes" (p.1277).

We also found evidence to suggest that reporting of relative p -values blurs the distinction between p and alpha. Some authors reported what were clearly intended to be exact p -values in a relative form (e.g. $p < .0732$), and others reported p in an exact form when clearly a relative value was intended (e.g. $z = 3.21$, $p = .05$).

Use of the term ‘level of confidence’ to refer to p -value (Practice 1) was short-lived, appearing only in 1955 and 1970. Chandler (1957) highlighted in the Psychological Bulletin the need for careful use of precise statistical terminology. He concluded:

The admixing of the concepts of confidence and significance has become so prevalent in the psychological literature that one typically reads statements ... indicating that certain experimental results were significant, at say, the 5% “level of confidence”.

It may be that this confusion arises from the fact that one can utilize a confidence interval as a significance test ... and in doing so may hastily, but incorrectly, conclude that there is no difference between the two concepts.
(p.430)

Chandler’s (1957) clarification may have been effective in reducing the use of this inaccurate terminology for NHST, but his discussion of confidence intervals appears to have had no impact on their uptake; we discuss confidence intervals below.

Further, over many years and still in 1999, authors, manuscript reviewers and JAP editors ignored some APA recommendations, without apparent explanation. Failure to specify a priori alpha violates APA (1994) recommendations, as does the apparent infrequent consideration of power. Almost all authors fail to follow the requirement to report p -values in abstracts which appeared first in the 1967 revision (APA, 1967) and in every Manual since. The more general conclusion is that APA Manual recommendations about statistical matters have not been an effective way to shape JAP publishing practices.

Language to describe statistical significance

Cohen (1994) stated

All psychologists know that statistically significant does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results section studded with asterisks implicitly becomes in the Discussion section highly significant or very highly significant, important, big! (p.1001, emphasis in the original)

We found widespread use of ambiguous language to describe ss and sns results (Figures 3 & 4); often the context suggested that ‘significant’ was intended to mean ‘important’. In our pilot work, we attempted to identify results that were described as both ‘statistically significant’ and, ambiguously, as ‘significant’. This coding proved problematic, but we noted many examples of a finding being described unambiguously and then ambiguously. For example, in a results section, the authors state: “The influence of costume on hiring recommendations was both positive and significant at the .05 level, with an F value of 3.30.” (our emphasis) However in the abstract of the same paper, the authors state: “... masculinity of the female applicant’s dress had a significant effect on interviewers’ selection decisions.” (our emphasis) In this example, although the term ‘statistical significance’ is not used in the statement of results, it meets our criteria for unambiguous statistical significance because the term ‘significant’ is clearly linked to the statistical test (Practice 5). Our criteria for statistical significance is less strict than the criteria used in some previous studies (e.g. Thompson & Snyder, 1997). In spite of this difference, we observed many ambiguous terms, and note that use of just the precise terminology, consistently through the whole paper, remains at a relatively low level in JAP papers (Figures 3 & 4). BJP papers in 1999 more often use just the unambiguous terms, but even here there is considerable room for improvement (Figures 3 & 4).

Reporting of *ss* and *sns* results

JAP reporting practices differ for *ss* and *sns* outcomes (Figures 3 & 4). Use of the ambiguous ‘non-significant’ alone is considerably more frequent than use of ambiguous ‘significant’ alone, especially in 1985 and 1999. Further, these data understate the difference, because *sns* results are often reported in a brief, shorthand way, with no textual reference to ‘significance’. For example, “There were no gender differences ($p > .05$)”. Overall, *sns* findings are given much less space and emphasis than *ss* findings.

Figure 5 may suggest good news in that the proportion of papers with one or more inferential statistics missing decreases markedly from 1955 to 1999. However, our estimates of the rates of missing statistics are certainly underestimates. We coded for missing inferential statistics (Practice 7) only when the unambiguous term ‘statistical (non) significance’ is used. When the ambiguous term is used, often the NHST is unreported; in particular, no inferential statistic is reported. Ambiguous terms are often used as a shorthand way to report a seemingly unimportant analysis.

Interpreting statistically non-significant results

Clearly, there are problems with authors’ interpretation of *sns* results. An average of 38% of all *sns* papers we studied included examples of accepting the null hypothesis (Figure 6); the practice was about as common in both journals in 1999 as for JAP in 1940. For reliability of coding, our criteria were relatively restrictive (see Practice 9 in Method). We noticed many other examples of accepting a null hypothesis that fell outside our criteria. For example, many papers reported a *sns* finding in the results section that was later, in the discussion, described as indicating that there were no differences. Our simple criteria required directly linked statements and so did not count such cases. Additionally, acceptance of a null hypothesis almost

never appeared in the context of a discussion of statistical power. The serious fallacy of accepting the null hypothesis is our most direct indicator of NHST problems.

Clearly, deep misconceptions persist.

Discounting of sns results (Practice 8) includes cases described as marginally significant or approaching significance, when p-values did not meet the standard for ss used elsewhere in a paper. We observed a rise in the proportion of sns papers showing this practice from an average of 14% (1940 to 1985) to 27% for JAP in 1999, even higher for BJP (Figure 6). Anecdotally, we noted that often p-values close to but greater than .05 were the only p-values reported exactly.

The discounting of sns results might be seen as a positive move away from routine reliance on the .05 level of significance as a dichotomous decision criterion. However Thompson (1993), and Thompson and Snyder (1997) argued that use of language such as ‘approaching significance’ is inappropriate and ambiguous. It is doubtful that the discounting we observed represents an informed interpretation of NHST results. Typically discounting occurs only when results ‘just fail’ a standard for statistical significance. We noted very few examples of researchers discounting ss results that ‘just meet’ a standard for statistical significance.

Our analyses of NHST reporting practices suggest a continued routine approach to statistical inference, and ongoing problems of interpretation. Psychologists continue to use .05 (although they also use .01 or .001) as a dichotomous decision criterion, and they present ss and sns findings differently. Further, researchers continue to use imprecise, ambiguous language to describe their findings.

Reform practices: statistical power

We observed a modest increase over time in the proportion of papers mentioning, however tangentially, statistical power or the role of sample size (Figure

7): from 13% of JAP papers in 1955-1970 to more than 40% in 1985-1999; the figure for BJP (1999) is 27%. In addition there are signs that authors in recent years are more explicit when they do mention issues relating to power. Here are two examples from 1940 and two from 1999:

1940:

The difference in per cent between those who received A and D ratings is not statistically reliable because of a limited number of subjects.

... the critical ratio was found to be 0.64; however, it will be noted that one of these groups has an N of only 16 drivers.

1999:

Our sample sizes for the undergraduate and graduate groups were large enough to allow use to report the results for each group separately (for $\eta^2 > .05$, power $> .75$).

Because of the large sample size of the study, many of these comparisons reached statistical significance. However, the absolute magnitude of these differences ... was typically not large.

However these results can give little comfort to reformers: Our Practice 10 had broad criteria and coded any recognition of sample size or power. Even so, in the sixty 1999 papers we studied from the two journals, even though almost all articles report sns results, fully two-thirds make no mention of power or the role of sample size, and only four indicate that power was calculated. Recalling also our observation that the null hypothesis is often accepted without consideration of power, it is clear that the APA (1994) recommendation to take power seriously has had little effect on practice.

Reform practices: variability and effect size

We found visual representations of variability in results in almost no JAP papers, and in just six papers (20%) in our BJP sample. We did not code for reports of effect size, but we noted very few examples of effect sizes being given a substantive interpretation. Our findings are consistent with those of others who have concluded that researchers often do not give the statistics or representations that are most useful for making substantive interpretations (Kirk, 1996; Thompson & Snyder, 1997, 1998) and for future meta-analyses.

Reform practices: confidence intervals

Only four of our JAP papers and one BJP paper report confidence intervals. The five examples are worth describing in some detail. In the sole BJP example the upper and lower confidence interval bounds are reported in a table but are not used to assist interpretation; NHST is carried out separately, based on the reported exact p -values. In a 1955 JAP paper, 90% confidence intervals appear in a table but are not even referred to in the text. In a 1985 JAP paper, a confidence interval is reported but not directly interpreted other than as a surrogate for a NHST: “The 95% confidence interval about this estimate excluded zero ($.02 < \rho < .35$) indicating that white raters assigned significantly higher ratings to white ratees than to black ratees.” In the third JAP example (1999) the dependent variable is psychological confidence, with negative scores indicating underconfidence and positive scores overconfidence: “In the lower group, there was slight underconfidence ($-.01$, 95% CI = $-.09$ [to] $.08$).” The mean is interpreted but the confidence interval is ignored as is the fact that the center of the confidence interval is close to zero.

After these four rather dismal examples, the final example (JAP, 1985) illustrates how confidence interval information can be substantively interpreted:

... the standard error of measurement [of the rating scores] was ± 40 points. Hence the 95% confidence interval range of 160 points encompassed four possible classification assignments. ... In a comparable worth context ... the rating score must serve as the primary criterion of relative worth; error variance of this magnitude is likely to be unacceptable.

These examples suggest that confidence intervals were not well understood even by most of the very few authors who report them. Only in the last example does the author make a substantive interpretation of the information provided by the interval.

Implications for reform

Our data are consistent with the findings of other studies: Distressingly little has changed over the last half century in the way psychologists report statistical inference. We will now consider the implications for reform, discussing first how some of the poor practices we identified might be improved, then general policy.

Dichotomous decision making. Our first general conclusion is that NHST is practised largely as dichotomous decision making that tends to keep researchers' attention on p -values and statistical significance, and away from substantive interpretation of findings. Encouragement of substantive interpretation is fundamental for reformers, including the TFSI. It must surely be at the center of all research, and statistical inference must support not hinder it. Kirk (1996), too, advocated careful judgement about the practical significance of findings:

It is a curious anomaly that researchers are trusted to make a variety of complex decisions in the design and execution of an experiment, but in the name of objectivity, they are not expected or encouraged to decide whether data are practically significant. (Kirk, p.755)

Many reformers have focussed on confidence intervals as the key alternative to NHST (see Harlow et al., 1997, for many examples). Here we will discuss how confidence intervals can avoid many of the problems of reporting practice that we found with NHST and, importantly, how they can support substantive interpretation.

A confidence interval presents an estimate of the true effect and its precision; This alone should encourage substantive interpretation. It comprises the best estimate of the true effect—for example the mean, usually at the center of the confidence interval—and an indication of the accuracy of that estimate: the width of the interval. Understanding and interpretation should be easy because this information is presented in the original measurement units. This information is also important because it is what is needed for future meta-analyses.

We found differences in reporting standards for ss and sns results, with shorthand reporting styles used more often for sns findings. Had researchers reported confidence intervals and focussed on substantive interpretation, it is likely the full range of results would have been reported and considered more equitably.

Statistical non-significance. We found frequent acceptance of a null hypothesis with little attention paid to statistical power; power estimates are almost never reported. When a conclusion of no practical difference or effect is being considered, a confidence interval may be especially helpful: It is a range of plausible values for the true effect, and so substantive interpretation of effects near the bounds is a good way of thinking about both precision and what conclusion is justifiable. A very narrow confidence interval that covers zero (or the parameter value that indexes no effect) suggests there is no practical effect. A wide confidence interval signals imprecision and low power, so even if it contains zero we should not conclude there is no true effect. Of course assessment of an interval as wide or narrow relies on researcher

judgment. If authors had examined confidence intervals many of the instances we noted of accepting the null hypothesis would surely have disappeared, and cases that remained would have been better justified.

This advantage of confidence intervals can also be exploited before running the experiment. Examination of a priori confidence interval width for various plausible values of effect size, population variance and sample size gives a good basis for making sound decisions about experimental design, including sample size. This practice may become accepted as more informative than using a priori estimates of power. In particular it can alert the researcher to the range of substantive interpretations that might result from a low power study.

Considering confidence intervals rather than NHST can help identify a plausible range of statistical inferences that might be made, and hence a range of plausible substantive interpretations of study results. This move away from NHST should reduce the temptations to rely on a dichotomous decision criterion, to use ambiguous language in describing statistical results, or to identify statistical significance with substantive importance.

Confidence interval issues. Confidence intervals undoubtedly have appeal but there is little evidence to support many of the claims made for them. Our conclusion above was that they are not well understood even by most of the few authors in our study who report them. We note further that the interpretation of a single confidence interval is not straightforward, and depends on an individual's definition of probability (Grayson et al., 1997; Reichardt & Gollob, 1997). These issues must raise at least some caution about claims (e.g. Hunter, 1997) that confidence intervals are easier to learn and understand than NHST.

A further question concerns the relation with NHST: Confidence intervals allow conduct of NHST simply by observing whether the interval covers the null hypothesized value. Since the focus is on points on the original measurement scale, not p -values, any conclusion about statistical (non)significance should be situated clearly in an appreciation of the distance between the observed mean and the null hypothesized value, in relation to the precision. The confidence interval approach to NHST may thus lead to more reasonable conclusions. There are however two potential problems: The first is that some people may use confidence intervals simply as a different way to do NHST. For these people, doing so may lessen the emphasis needed on effect size, precision and substantive interpretation. Indeed some reformers (e.g. Hunter, 1997) advocate use of confidence intervals without any reference to NHST. Second, a confidence interval approach to NHST that notes simply whether or not the interval includes the null hypothesized value may divert attention from exact p -values and thus be a backward step.

Research is needed on these and other confidence interval questions, and the outcomes of this research are likely to give important guidance to reform. We take up these issues more fully in Finch et al. (2000).

Promoting reform

The main lesson of our findings is that by 1999 precious little reform of JAP statistics reporting practices had happened. We did not code for reporting of effect size measures, and others have reported that this may be more frequent in recent years, but even when reported they are often not used to support substantive interpretation (Thompson & Snyder, 1997, 1998). In our data there is a modest trend for awareness of sample size and power to be more frequent in recent years, but even here the proportion has simply returned to roughly the level of 1940. Further, fully

two-thirds of the 1999 papers make absolutely no mention of such issues. Our main conclusion remains: many important aspects of inference practices and reporting were the same in 1999 as half a century earlier. Judged by the papers we analyzed, the cogent, sustained efforts of the reformers have been a dismal failure.

It is worth emphasizing again that we are not referring to mere inconsequential differences in fashion. Psychology's routine NHST practices, as the meta-analysts in particular have demonstrated, regularly lead to seriously wrong conclusions and waste of large amounts of research effort. Achieving reform actually does matter and will make a difference.

Why has reform proceeded further in some other disciplines, including medicine, than in psychology? The American Medical Association's "Uniform requirements for Manuscripts submitted to Biomedical Journals", sets out simple requirements for several important reform practices, including routine use of confidence intervals and avoidance of ambiguity in use of statistical terms (International Committee of Medical Journal Editors, 1997). Note how radically different it is from current APA guidelines. What has happened in psychology was not inevitable.

We leave to historians and sociologists of science the fascinating and important question of why psychology has persisted for so long with poor statistical practice.

Two aspects of our findings may have consequences for future reform policy. First, we found clear evidence that publishing a recommendation in the APA Manual did not necessarily shape JAP reporting practice. We do not know why editors and authors follow some of the Manual's statistical guidelines but, without explanation, largely ignore others; this is a further question for sociologists of science. In some important ways the guidelines are ahead of practice, in the sense of being more reform-oriented. For example, despite the Manual, effect sizes are in many cases not

reported and power is seldom considered seriously. Our conclusion is that a range of APA's statistical guidelines have been ineffective in shaping practice. Reform will not be achieved simply by revising the APA Manual to recommend the full reform agenda, important though such revisions would be.

A second issue is the central role of journal editors. In relation to JAP editors we can say that they made a number of brief editorial statements that showed awareness of some reform issues. We do not know whether they adopted any policies to promote reform, but we found very little sign in JAP's published papers of progress towards reform, even in 1999.

Shrout (1997) reported one apparently successful case of an editor in epidemiology implementing a “virtual ban” (p. 1) on NHST. Individual editors of JAP, as of other psychology journals, could with advantage have adopted such a policy, although it is uncertain how substantial a change any single editor could have achieved. Editors of many journals acting in concert would be more likely to achieve substantial change.

Conclusion

Most broadly, psychologists in every research role must share responsibility for the failure of reform: writers of statistics texts and software, statistics teachers, researchers themselves, journal editors and manuscript reviewers, and participants in APA and other policy-making bodies. Correspondingly, all need to take responsibility for promoting change.

The TFSI report appeared too recently to influence reporting practice in the papers we studied. It is a welcome development and we strongly support almost all of its recommendations. However it remains unclear what steps will be most effective in now bringing about reform; our conclusion is that more is needed than simply a

revised set of APA guidelines and statements from journal editors. Efforts on a broader front will be needed to disturb half a century of stagnation in psychologists' reporting practices of statistical inference.

Figure captions

Figure 1

(Practice 2) Proportion of papers reporting exact and relative p-values (standard error bars shown)

Figure 2

(Practices 3 and 4) Proportion of papers reporting a priori alpha and more than one implicit alpha level (standard error bars shown)

Figure 3

(Practices 5 and 6), Proportion of papers using unambiguous ‘statistical significance’, ambiguous ‘significance’ or both terms when reporting ss results (standard error bars shown)

Figure 4

(Practices 5 and 6), Proportion of papers using unambiguous ‘statistical non-significance’, ambiguous ‘non-significance’ or both terms when reporting sns results (standard error bars shown)

Figure 5

(Practice 7) Proportion of papers failing to report a test statistic when reporting ss or sns results (standard error bars shown)

Figure 6

(Practices 8 and 9) Proportion of sns papers discounting a statistically non-significant result or accepting a null hypothesis (standard error bars shown)

Figure 7

(Practice 10) Proportion of papers in which authors report awareness of sample size or in which power calculations are mentioned (standard error bars shown)

References

- American Psychological Association Council of Editors (1952). Publication Manual of the American Psychological Association. Psychological Bulletin, 49, 389-450.
- American Psychological Association (1957). Publication Manual of the American Psychological Association, 1957 Revision. Washington DC: Author.
- American Psychological Association (1967). Publication Manual of the American Psychological Association, 1967 Revision. Washington DC: Author.
- American Psychological Association (1974). Publication Manual of the American Psychological Association (2nd edition). Washington DC: Author.
- American Psychological Association (1983). Publication Manual of the American Psychological Association (3rd edition). Washington DC: Author.
- American Psychological Association (1994). Publication Manual of the American Psychological Association (4th edition). Washington DC: Author.
- Bakan, D. (1967). On method. San Francisco: Jossey-Bass.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526-542.
- Bobko, P. (1995). Editorial. Journal of Applied Psychology, 80, 3-5.
- Boring, E.G. (1919). Mathematical versus statistical significance. Psychological Bulletin, 16, 335-338.
- Campbell, J.P. (1982). Editorial: Some remarks from the Outgoing Editor. Journal of Applied Psychology, 67, 691-700.
- 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.
- Chase, L.J., & Chase, R.B. (1976). A statistical power analysis of Applied Psychological Research. Journal of Applied Psychology, 61, 234-237.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and 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. (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.
- Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, 5, 75-98.
- Finch, S., Thomason, N., & Cumming, G. (2000) Past and future APA guidelines for statistical practice. Manuscript in preparation.
- 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 behavioural sciences: Methodological issues. Hillsdale, NJ: Erlbaum.
- Gigerenzer, G., & Murray, D. (1987). Cognition as intuitive statistics. Hillsdale, NJ: Erlbaum.
- Grayson, D., Pattison, P., & Robins, G. (1997). Evidence, Inference and the “Rejection” of the Significance Test. Australian Journal of Psychology, 49(2), 64-70.
- Guion, R.M. (1983). Editorial: Comments from the New Editor. Journal of Applied Psychology, 68, 547-551.

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

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

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 & Psychology, 7, 545-554.

Hunter, J.E. (1997). Needed: A ban on the significance test. Psychological Science, 8, 3-7.

Hunter, J., & Schmidt, F. (1990). Methods of meta-analysis. Newbury Park, CA: Sage.

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

John, I.D. (1992). Statistics as rhetoric in psychology. Australian Psychologist, 27, 144-149.

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

Loftus, G.R. (1993). 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. (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.T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70, 151-159.

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.

Muchinsky, P.M. (1979). Some changes in the characteristics of articles published in the Journal of Applied Psychology over the past 20 years. Journal of Applied Psychology, 64, 455-459.

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

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

Pollard, P. (1993). How significant is “significance”? In G. Keren & C. Lewis (Ed.s), A handbook for data analysis in the behavioural sciences: Methodological issues. Hillsdale, NJ: Erlbaum.

Pollard, P., & Richardson, J. (1987). On the probability of making Type I errors. Psychological Bulletin, 102, 159-163.

Reichardt, C.S., & Gollob, H.F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. Harlow, S. Muliak, & J. Steiger (Eds.), What if there were no significance tests? (pp. 259-284) Mahwah, NJ: Lawrence Erlbaum.

Rosnow, R.L., & Rosenthal, R. (1989) Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rossi, J.S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. Harlow, S. Muliak, & J.

Steiger (Eds.), What if there were no significance tests? (pp. 175-197) Mahwah, NJ: Lawrence Erlbaum.

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.

Schmitt, N. (1989). Editorial. Journal of Applied Psychology, *74*, 843-845.

Shrout, P. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. Psychological Science, *8*, 1-2.

Thompson, B. (1993). Foreword. The Journal of Experimental Education, *61*(4), 285-286.

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., & Snyder, P.A. (1997). Statistical significance testing practices in The Journal of Experimental Education. The Journal of Experimental Education, *66*, 75-83.

Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling & Development research articles. Journal of Counseling & Development, *76*, 436-441.

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

Uniform requirements for Manuscripts submitted to Biomedical Journals. (1997).
Journal of the American Medical Association, 277, 927-934.

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

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.

Author notes

The authors acknowledge the support of the Australian Research Council. Bruce Thompson provided helpful comments on a draft. Joanna Leeman, Brooke Macpherson and Andrew McClure assisted with the pilot work for this study.

Correspondence about this article should be addressed to Sue Finch, Department of Mathematics & Statistics, The University of Melbourne, Australia, 3010. Email: s.finch@ms.unimelb.edu.au