

Running Head: STATISTICAL REFORM LESSONS FROM MEDICINE

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119-126

© American Psychological Society. This article may not exactly replicate the final version published in the journal. It is not the copy of record. The definitive version is available at www.blackwell-synergy.com.

Editors can Lead Researchers to Confidence Intervals, but Can't Make Them Think
Statistical Reform Lessons From Medicine

Fiona Fidler^{1,2}, Neil Thomason², Geoff Cumming¹, Sue Finch¹ and Joanna Leeman¹

1 = La Trobe University, Melbourne, Australia

2 = The University of Melbourne, Melbourne, Australia

Author contact:

Fiona Fidler

Department of History and Philosophy of Science
University of Melbourne, Victoria, Australia 3010

Phone: +61 3 8344 5579

Email: fidlerfm@unimelb.edu.au

Abstract

Since the mid-1980s, confidence intervals (CIs) have been standard in medical journals. The authors sought lessons for psychology from medicine's experience of statistical reform by investigating two attempts by Kenneth Rothman to change statistical practices. They examined 594 *American Journal of Public Health (AJPH)* and 110 *Epidemiology* articles. Rothman's editorial instruction to report CIs and not *p* values was largely effective: In *AJPH* sole reliance on *p* values dropped from 63% to 5%, and CI reporting rose from 10% to 54%; *Epidemiology* showed even stronger compliance. However compliance was superficial: Very few authors referred to CIs when discussing results. These results support what other research has indicated: Editorial policy alone is not a sufficient statistical reform mechanism. Achieving substantial, desirable change will entail considerable guidance for full use of CIs and appropriate effect size measures. This will require study of researchers' understanding of CIs, improved education, and development of empirically-justified recommendations for improved statistical practice.

Statistical significance testing (or null hypothesis significance testing, NHST) has been criticised in many disciplines, not only psychology (Anderson, Burnham, & Thompson, 2000; Morrison & Henkle, 1970). Yet within psychology there has rarely been mention of such criticisms, or of reform efforts in ecology, medicine and other disciplines. An exception was Shrout's (1997) discussion of an editorial ban on NHST in the *American Journal of Public*

Health (AJPH) that, he claimed, led to dramatically improved statistical practices. In psychology, by contrast, journal editorial policy has been largely ineffective in improving statistical practices. In this paper we investigate reform efforts in epidemiology and seek lessons for psychology. We focus on confidence intervals (CIs) and effect size reporting, because these are highly desirable and now recommended in psychology (American Psychological Association [APA], 2001).

Little Interdisciplinary Discussion of NHST

In their 1970 anthology, *The Significance Test Controversy*, Morrison and Henkle noted that there had been little exchange over NHST issues, even between closely related disciplines. Their book contained chapters from both psychologists and sociologists. Until that time, scrutiny of NHST in the two disciplines had been “parallel but quite independent.” (Morrison & Henkel, p. 182) Unfortunately, there has been little exchange since.

In the mid-1990s the American Psychological Association and the American Psychological Society held symposia to discuss banning NHST from psychology journals. Resulting articles taking various perspectives were published in *Psychological Science* in January 1997. Shrout’s (1997) introduction to these articles included a brief discussion of the *AJPH* story and was the only attempt we know of to draw attention to a ban outside psychology.

In 1996 the APA Task Force on Statistical Inference (TFSI) was appointed to investigate the ban proposal. The *AJPH* case study was, surprisingly, never systematically investigated by the TFSI. For a discipline that claims to be empirical, psychology has been strangely uninterested in evidence relating to statistical reform.

Statistical Reform in Psychology: CIs are Important but Rarely Used

For decades many advocates of statistical reform in psychology have recommended CIs as alternative (or at least supplementary) to p values and the APA *Publication Manual* now calls them “the best reporting strategy.” (APA, 2001, p.22). While it is too early to assess the impact of the new APA recommendation, previous decades of encouraging researchers to report CIs have had little effect. They remain relatively little-used (Finch, Cumming, & Thomason, 2001; Kieffer, Reese, & Thompson, 2001).

Researchers rarely see CIs reported, so some may not understand why they are important; researchers may also have the misconception that CIs are equivalent to NHST. CIs can indeed be used to perform NHST, by noting whether the null value is within the interval. Unlike p values, however, they also provide information on precision: CI width is a guide to this. Furthermore, effect sizes are important because they are the primary outcome of research and are needed for meta-analysis, and CIs are estimates of effect size. Unlike NHST, CIs “provide information on both location and precision.” (APA, 2001, p. 22).

Editorial Policy and Statistical Reform in Psychology

In 1993, editor-elect of *Memory & Cognition*, Geoffrey Loftus, wrote that he intended “to decrease the overwhelming reliance on hypothesis testing” (Loftus, 1993, p.3). He encouraged the use of figures with error bars, pointing out that “more often than not, inspection of such a figure will immediately obviate the necessity of any hypothesis-testing procedures” (p.3). Loftus recently explained: “...the goal is simple—to have data be as comprehensible to its consumers as is possible. And I believe that confidence intervals serve that goal and I believe hypothesis testing doesn’t serve that goal” (G. Loftus, personal communication, August 2001).

Finch, Cumming, Williams, et al (2004) studied the Loftus reform attempt. They found that during Loftus' term as editor the proportion of articles reporting error bars (whether CIs or standard error bars) did increase—from 7% under the previous editor to 41%. However, after Loftus departed the proportion dropped to 24%.

Loftus and his editorial board worked hard at enforcing his policy, calculating many error bars for authors. However, even with this assistance, at most half the articles adopted the recommendations. Furthermore, NHST was reported and used as the basis for interpretation in almost all empirical articles; even when CIs were reported they were seldom used for interpretation.

In 1997 Phillip Kendall stated as policy that authors submitting to the *Journal of Consulting and Clinical Psychology (JCCP)* should report effect sizes and discuss clinical significance. Fidler, Cumming, Thomason et al (2005) investigated this *JCCP* reform attempt and found that frequency and types of effect size reporting have shown little change over the last decade, despite the 1997 editorial. Similarly, 38% of 1996 articles discussed clinical significance, distinguished from statistical significance; in 2000-01 still only 38% of articles did this.

Vacha-Haase et al (2000) reviewed 10 independent studies of reporting practices in 23 psychology and education journals. They reported that “effect sizes have been found to be reported in *between roughly 10 percent ... and 50 percent of articles...* notwithstanding either historical admonitions or the 1994 manual's ‘encouragement’” (p. 419, emphasis in original).

Similarly, Finch, Cumming and Thomason (2001) found little evidence of reform in the statistical reporting practices of the *Journal of Applied Psychology (JAP)* over the last 60 years. Of 150 *JAP* articles, only 4 reported CIs. Thirty *British Journal of Psychology* articles from 1999 were coded for comparison; only 1 reported a CI. (Finch, Cumming & Thomason, 2001).

Together these surveys suggest that lasting reform will require more than piecemeal editorial policy change, even if initiated by committed and energetic editors.

Statistical Reform in Medicine: CIs are Reported but not Interpreted

At first glance, editorial policy changes in medical journals seem effective: ShROUT stated that in the *AJPH* postban period “Confidence intervals...are evident in most articles” (p.1). But as we saw with Loftus' reform, simply reporting CIs does not guarantee that they will be used to interpret the data. Savitz, Tolo and Poole (1994) explained that even though 70% of articles in the *American Journal of Epidemiology (AJE)* reported CIs “inferences are made regarding statistical significance tests, often based on the location of the null value with respect to the bounds of the confidence interval” (p.1051). That is, CIs were simply used to do NHST.

By 1984, when Rothman became assistant editor at the *AJPH*, discussions about statistical reform in medicine were well advanced. In 1977 the *New England Journal of Medicine (NEJM)* reviewed its statistical reporting problems. Soon after the *Journal of the American Medical Association (JAMA)* and *Circulation Research* made similar efforts to improve their reporting practices (Rennie, 1978). By 1986, the *British Medical Journal (BMJ)* had a policy encouraging CIs (Langman, 1986).

In 1988, the International Committee of Medical Journal Editors (ICMJE) revised their “Uniform Requirements for Manuscripts Submitted to Biomedical Journals”:

When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of *p* values, which fail to convey important quantitative information...(p.260)

The “Uniform Requirements” were published in the *Annals of Internal Medicine* (ICMJE, 1988a) and the *BMJ* (ICMJE, 1988b). Over 300 medical and biomedical journals notified ICMJE of their willingness to comply with the manuscript guidelines, including *AJPH*. In 1989, *AJPH* began quoting the ICMJE guidelines (above) in each issue’s “Instructions to Authors”.

Rothman did not, as ShROUT (1997) suggested, *ban* NHST outright: There was no such policy stated by the journal. However, his revise and submit letters were unambiguous:

All references to statistical hypothesis testing and statistical significance should be removed from the papers. I ask that you delete p values as well as comments about statistical significance. If you do not agree with my standards (concerning the inappropriateness of significance tests) you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere (ShROUT, 1997, p.1, citing Fleiss, 1986, citing Rothman).

As assistant editor Rothman saw only approximately 25% of the manuscripts. Alfred Yankauer, the editor, eventually “came around” to Rothman’s views, though Rothman suspects he was not as persistent in requesting such revisions (K. Rothman, personal communication, September 2001). Still, the message was certainly stronger than many of the *encouragements* that have been made in psychology. Here Rothman describes his goal:

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

Rothman’s letters created some controversy. In 1986, *AJPH* published Fleiss’ defence of NHST, where he argued that “An insidious message is being sent to researchers in epidemiology that tests of significance are invalid and have no place in their research” and that significance tests “have a valid role to play at each important step in the analysis of epidemiological data” (p.559). From a review of the medical literature, it seems there was little explicit support of Fleiss’ view; other defences of NHST are difficult to find.

Other criticisms concerned the process used to implement the change, rather than the change itself. ShROUT, for example, agreed with a shift from NHST to CIs, but was displeased that Rothman’s “proposal was not discussed in forums like this [the symposia that led to the special section in *Psychological Science*]” (1997, p.1). However, by the time Rothman started with *AJPH*, there had been substantial criticism of NHST and discussion of CIs in the medical literature (e.g. Altman, 1980; Mainland, 1984). Rothman himself had published views on this topic before his *AJPH* appointment. For example, in 1978 he wrote the *NEJM* editorial “A show of confidence”.

In 1991 (when a new editor replaced Yankauer) the entire section on statistical reporting was cut from *AJPH*’s “Instructions to Authors”. Recent issues have only a general reference to the ICMJE guidelines; they do not include any recommendations specific to analysis or statistical reporting.

Rothman founded *Epidemiology* in 1990. As editor, he instituted an even stricter policy on NHST. Rothman’s fellow editors were sympathetic to his views and, he believes, at least as strict as he was in enforcing the policy (K. Rothman, personal communication, September 2001).

Effect Sizes

Twenty journals in psychology and education now require effect size reporting (Thompson, 2002). Some particularly advocate *standardised* effect sizes (e.g., Cohen’s d), which often facilitate meta-analysis (Hunter & Schmidt, 1990; Thompson, 1994, 1996; Wilkinson et al,

1999). Kirk (1996), for example, identified and explained over 40 different measures of effect size. Because effect size measures are, alongside CI use, at the heart of reform goals in psychology, we also examined effect size practices.

Method

We surveyed *AJPH* articles published before, during and after Rothman's policy, before and after the ICMJE regulations, and before and after changes to the "Instructions to Authors". We coded 594 *AJPH* articles, published in selected years between 1982 and 2000 (Table 1). From *Epidemiology* we coded 40 articles published in 1990, the year Rothman founded the journal, and 70 from 2000, his final year as editor. We recorded a practice (e.g., CI use) as present if an article contained at least one instance of the practice; we did not count any further instances.

Table 1.

Publication years chosen for coding of American Journal of Public Health (AJPH) articles, number of articles coded, and reasons for interest in those years

Publication year	<i>N</i> (number of articles coded)	Reason for choosing year, and some possible influences on practices seen in that year
1982	67	Pre-Rothman
1986	98	Expected maximum influence of Rothman, whose term was 1984 to February 1987
1988	71	Immediate post-Rothman.
1989	72	Post-Rothman
1990	72	Post-Rothman. ICMJE recommendations (published 1988, referred to in <i>AJPH Instructions to authors</i> in 1989)
1993	72	New editor, and specific reference to ICMJE recommendations dropped from <i>Instructions to authors</i> in 1991.
1994	72	As 1993
2000	70	Recent practices

Items Coded

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

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

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

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

Reliability of Coding

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

Results

Statistical Significance Testing

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

Confidence Intervals

Of the 594 *AJPH* articles, 322 (54%) reported CIs; of 110 *Epidemiology* articles, 95 (86%) did so. Fully 268 (83% of 322) of these *AJPH* articles reported CIs in tables (usually very large ones); only 15 (5%) displayed error bars in figures. In *Epidemiology*, the corresponding frequencies were 81 (85% of 95) and 6 (6%).

Table 2 shows that fewer than 12% of *AJPH* articles with CIs interpreted them and that, despite fully 86% of articles in *Epidemiology* reporting CIs, interpretation was just as rare. Nor did the situation improve over time. For example, only 1 (of 40) 1990 *Epidemiology* paper referred to CI width; in 2000, only 3 (of 70 papers) did this.

Table 2.

Types and frequency of CI interpretation in AJPH and Epidemiology

	<i>AJPH</i> % of 322 (number)	<i>Epidemiology</i> % of 95 (number)
Any mention of CI limits	1.2 (4)	1.1 (1)
Any mention of CI width	2.2 (7)	4.2 (4)
Any mention of CI overlap	1.9 (6)	0
Any reference to null value	6.2 (20)	3.2 (3)
Any CI interpretation ^a	11.2 (36 ^a)	8.4 (8)

^aOne *AJPH* article had two interpretations.

NHST vs CIs (1982-2000)

Figure 1 shows that sole reliance on NHST or p values dropped dramatically during Rothman's term at *AJPH*, from 63% in 1982, to 6% in 1986-9. CI reporting increased from 10% before Rothman, to 54% in 1986 towards the end of his editorial service. As shown in Figure 1, the changes in percentages mentioned here are large by comparison with the 95% CI widths. In *Epidemiology*, CIs were even more common. Fully 94% of articles in 2000 reported them. P values were rare; there were none in 2000.

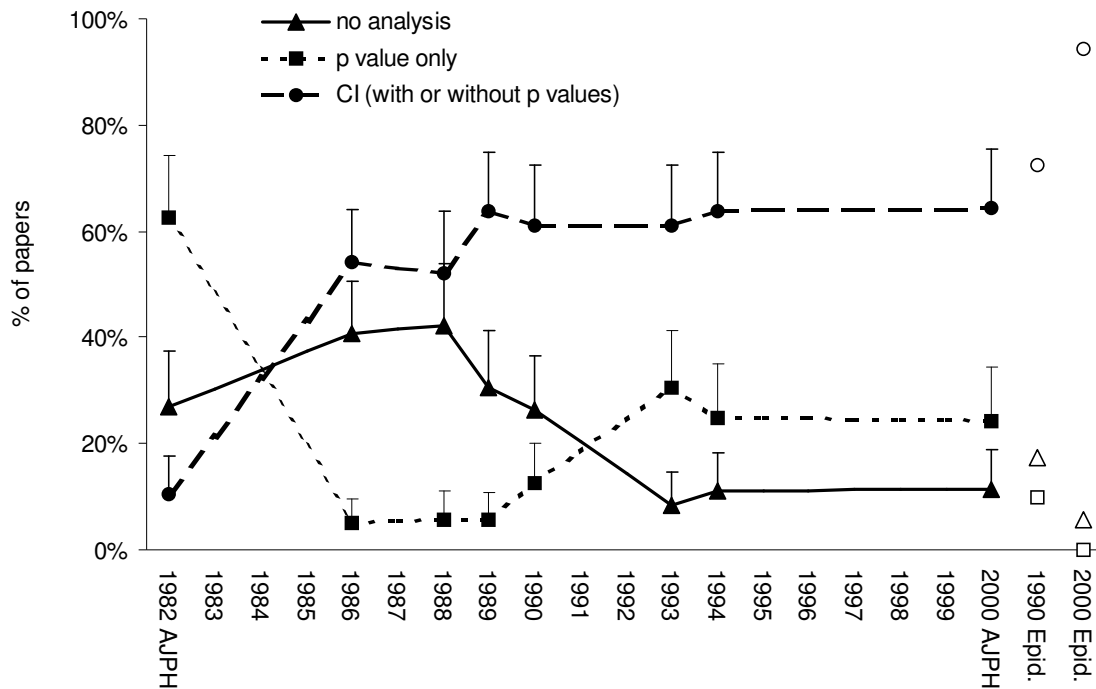


Figure 1. Percent of *AJPH* and *Epidemiology* (*Epi.*) articles reporting at least one p value, at least one CI, or no instances of either. Error bars are upper half 95% CIs. Open symbols for *Epi.* give the values for the corresponding closed symbols for *AJPH*.

Figure 1 also shows a concomitant increase from 1982 to 1986-8 in no-analysis, or non-inferential articles in *AJPH*. These articles often reported effect sizes (e.g., means, percentages, ratios), but included neither NHST results nor CIs. Some were legitimate descriptive studies of entire populations, and so did not require inference. Others, however, contained evidence that authors were doing *covert* significance testing. Occasionally there was explicit evidence for this, for example a footnote explaining that NHST, whilst not reported, had been conducted and readers were invited to contact authors for results. In other articles there was ambiguity: Although no NHST results were reported, discussions focused on "significant differences".

Figure 1 suggests the CI reporting in *AJPH* has been relatively stable since Rothman left. However, Figure 2 tells an interestingly different story. It shows that, while Rothman was at *AJPH*, CIs were commonly reported without p values. For some time after his departure in early 1987 this trend remained. By 1993, however, the number of articles reporting p values had increased dramatically. CIs continued to be reported, but from this point on were supplementing p values, rather than replacing them. For example, in 1990, 42% of articles reported only CIs and a further 19% reported both CIs and p values. In 1993, these figures were reversed: 13% reported

only CIs and 48% reported both CIs and p values. There was also an increase in sole reliance on p values (Figure 1): In 1990, less than 13% of articles relied only on p values; by 1993, this figure had more than doubled at just over 30%. This resurgence in p values followed the arrival of a new editor, and removal of the ICMJE's recommendations from *AJPH*'s "Instructions to Authors" in 1991.

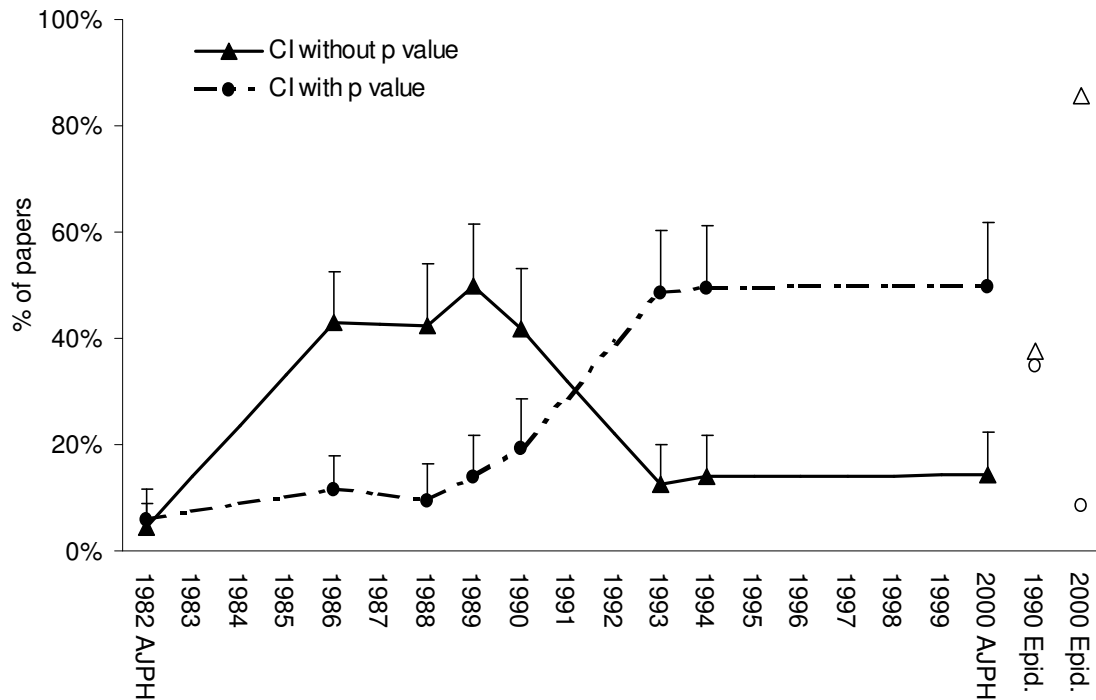


Figure 2. Percent of *AJPH* and *Epi.* articles that reported both p values and CIs, and articles that reported only CIs. Error bars are upper half 95% CIs. Open symbols for *Epi.* give the values for the corresponding closed symbols for *AJPH*.

Effect Sizes

Figure 3 shows that effect sizes were very often reported as percentages and proportions (*AJPH* 95%, *Epidemiology* 84%). Means and mean differences were common in *Epidemiology* (92%) but not so much in *AJPH* (38%). Odds ratios and relative risk values were reported in approximately half of the articles in both journals (*AJPH* 46%, *Epidemiology* 53%). Other units-free measures were reported only occasionally (e.g., in *AJPH* 13% of articles reported correlation coefficients, 9% R^2 values). There were no reports of effect sizes in standard deviation units (e.g., Cohen's d) in either journal. These rates were stable over the years surveyed.

Discussion

Our data, as summarized by Figures 1 and 2, suggest that Rothman's efforts at *AJPH* were initially effective by leading to a remarkable drop in NHST use and increase in CI reporting. Several years after his departure CI reporting remained high—as it had become in many other medical journals—but p values had again become common in *AJPH*. He was more consistently successful at *Epidemiology*.

In both journals, however, when CIs were reported they were rarely used to interpret results. This rather ominous finding holds even for the most recent years we surveyed. In

addition, in many *AJPH* and *Epidemiology* articles in which NHST and p values were not explicitly reported, there was evidence, or at least clear hints, that interpretation was based on unreported NHST.

Almost all articles reported some effect sizes, often percentages and ratios. Cohen's d and similar standardized effect size measures that are often needed in psychology for meta-analysis were not used.

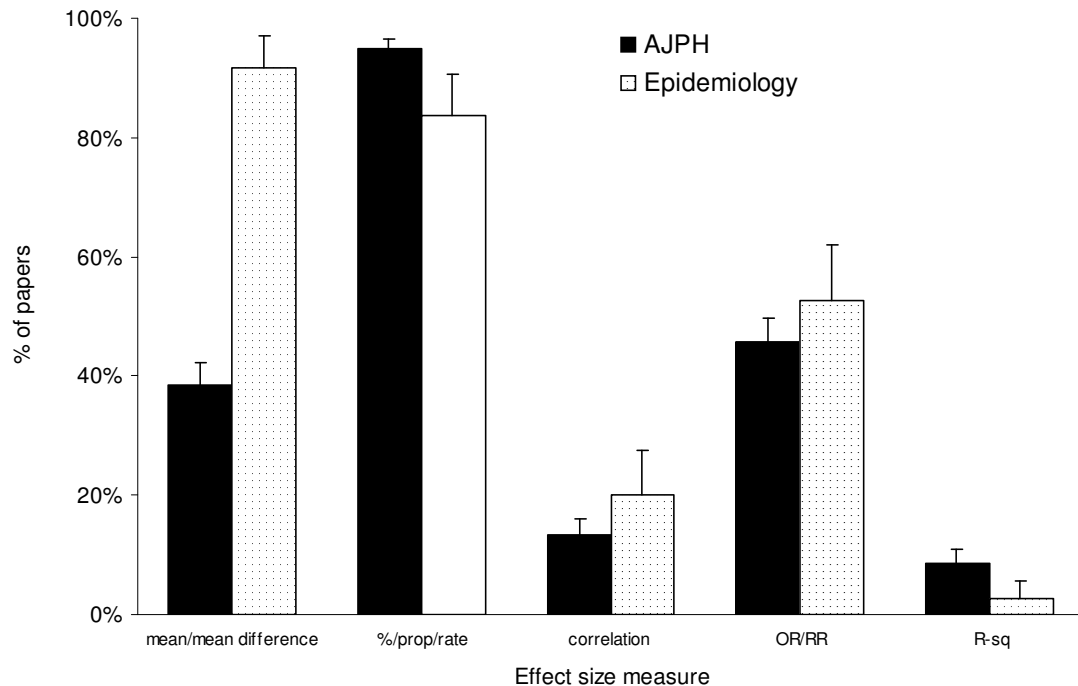


Figure 3. Percent of *AJPH* and *Epi.* reporting at least one of the listed effect size measures. Error bars are upper half 95% CIs. We did not find any instances of the following effect size measures: Cohen's d , η or η^2 , ω or ω^2 .

Both Rothman and Loftus achieved considerable change in some aspects of statistical reporting, but in both cases these changes did not persist beyond their editorial incumbency, and the more fundamental change of using CIs for interpretation, was largely not achieved. One interesting difference was that Rothman reported little resistance from authors, (K. Rothman, personal communication, September 2001) whereas Loftus encountered considerable resistance, and even himself calculated error bars for around 100 authors who failed or refused (!) to include them in their manuscripts. (G. Loftus, personal communication, August 2001)

Several differences between the contexts in which Rothman and Loftus worked should be noted. Medical researchers commonly consult with statisticians; in psychology this occurs less frequently. Also, at the time of Rothman's reforms, many major medical journals (including *NEJM* and *BMJ*) were reforming their statistics reporting policies, then in 1988 the ICMJE guidelines were revised. Loftus did not have analogous support from the APA—his reforms at *Memory & Cognition* started in 1993, but the APA *Publication Manual* did not recommend CIs until 2001. A change of editor and removal the ICMJE's guidelines (from *AJPH*'s "Instructions

to Authors”) in 1991 may also be important. Between 1991-1993 p value reporting increased dramatically: CIs, whilst still present, were rarely the sole reporting mechanism.

A Lesson for Psychology: Editors can Lead Researchers to Confidence Intervals but They Can't Make Them Think

The most striking lesson here is that even authors who presented CIs and not p values largely ignored CIs when interpreting their data. Savitz, Tolo and Poole (1994) reported that in the *AJE* “the most common practice was to provide confidence intervals in results tables and to emphasize statistical significance tests in result text” (p.1047). Our results show this continues. In *AJPH* and *Epidemiology* we found a large proportion of CIs in (usually) very large tables, but rarely in figures, which we believe would enhance their value (Cumming & Finch, 2005). Furthermore, even when, as at *Epidemiology*, editorial pressure is maintained and NHST is rare, authors still rely on terms like ‘significant difference’ to interpret results. Though CIs were reported in almost all *Epidemiology* articles, they seemed to have virtually no effect on the way authors interpreted their data. For example, less than 5% mentioned the CI width, despite this information on precision being one of the greatest advantages of CIs over NHST. The *Memory and Cognition* survey demonstrates that CI interpretation is also rare in psychology (Finch, Cumming, Williams et al, 2004).

Merely reporting CIs is an important first step: It allows the sagacious reader to better understand the data’s implications. But transcending superficial reporting changes will require more than the efforts of lone editors.

We fear that one reason that policy changes have been, and will continue to be, insufficient is because relevant knowledge is lacking. Little is known about how researchers think about CIs, or what misconceptions might be associated with their use. There is little empirically-based guidance about how to best present, interpret or discuss CIs. Although some reformers have claimed that CIs are easier to teach and less frequently misinterpreted than NHST (e.g., Hunter & Schmidt, 1997; Schmidt, 1996), these claims are supported only by anecdotal evidence. Also, it remains unclear the extent to which CIs can replace NHST, and the extent to which p values *should* be replaced (Cumming & Finch, 2001).

We recommend several lines of systematic investigation:

- empirical studies of how researchers presently interpret CIs (e.g., Belia, Fidler, Williams, & Cumming, 2005; Cumming, Williams, & Fidler, 2004);
- empirical and conceptual studies of how CIs can most effectively be presented and used to interpret research results (e.g., Cumming & Finch, 2005);
- empirical studies of how to teach CIs, and the consequences of giving CIs a foundational role in statistics education. For decades, defenders of NHST used the bromide of “better teaching” as their solution to NHST’s problems, but rarely, if ever, investigated what was needed to improve statistics education. Those advocating CIs should not fall into the same comfortable trap.

In addition, substantial revamping of the *APA Publication Manual* is needed. Although it now strongly advocates CIs, it gives no examples of how they should be presented, or advice on how to use them to interpret results. (Fidler, 2002; Finch, Thomason & Cumming, 2002)

Finally, as Morrison and Henkel (1970) advocated more than 30 years ago, psychology’s long tradition of ignoring statistics reform debates in other disciplines should end. Psychology needs all the help it can get. Despite remaining challenges, the enterprise of pioneering editors,

such as Rothman and Loftus, deserves considerable praise. Their work provides valuable and scarce empirically-based guidance as to how statistical reform might best proceed.

References

- Altman, D. G. (1980). Statistics and ethics in medical research: VI—presentation of results. *British Medical Journal*, *281*, 1542-1544.
- Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence and an alternative. *Journal of Wildlife Management*, *64*, 912-923.
- American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 530-572.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist*, *60*, 170-180.
- Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*, 299-311.
- Fidler, F. (2002). The 5th edition of the APA Publication Manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62*, 749-770.
- Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., & Schmitt, R. (2005). Evaluating the effectiveness of editorial policy to improve statistical practice: The case of the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology*, *73*, 136-143
- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181-210.
- Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of *Memory & Cognition*. *Behavior Research Methods, Instruments & Computers*, *36*, 312-324.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, *12*, 825-853.
- Fleiss, J.L. (1986) Significance tests do have a role in epidemiological research: Reaction to A. M. Walker. *American Journal of Public Health*, *76*, 559-60.
- Hunter, J., & Schmidt, F. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In L. Harlow, S. Mulaik & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-63). Mahwah, NJ: Erlbaum.
- Instructions to Authors. (1989). *American Journal of Public Health*, *79*(4), 525.
- International Committee of Medical Journal Editors. (1988a). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine*, *108*, 258-265.
- International Committee of Medical Journal Editors. (1988b). Uniform requirements for manuscripts submitted to biomedical journals. *British Medical Journal*, *296*, 401-408.

- Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, *69*, 280-309.
- Kirk, R. (1996) Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.
- Langman, M. J. S. (1986) Towards estimation and confidence intervals. *British Medical Journal*, *292*, 716.
- Loftus, G. R. (1993). Editorial comment. *Memory & Cognition*, *21*, 1-3.
- Mainland, D. (1984). Statistical rituals in clinical trials: Is there a cure? *British Medical Journal*, *288*, 341-343.
- Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. Chicago, IL: Aldine.
- Rennie, D. (1978). Vive la difference ($p < 0.05$). *New England Journal of Medicine*, *299*, 828-829
- Rothman, K. (1978). A show of confidence. *New England Journal of Medicine*, *299*, 1362-1363.
- Savitz, D. A., Tolo, K., & Poole, C. (1994). Statistical significance testing in the *American Journal of Epidemiology*, 1970-1990. *American Journal of Epidemiology*, *139*, 1047-1052.
- Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.
- 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. (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*, 26-30.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 24-31.
- 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. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Author note

The authors may be contacted by email: Fiona Fidler: fidlerfm@unimelb.edu.au, Neil Thomason: neilt@unimelb.edu.au, Geoff Cumming: G.Cumming@latrobe.edu.au. The authors thank Geoffrey Loftus for interviews that we quote from in this article and Kenneth Rothman for both his correspondence and comments on our manuscript. We also thank Patrick Shrout for early discussions about this project. This work was supported by the Australian Research Council.