

Running Head: STATISTICAL REPORTING IN *JCCP*

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.

© American Psychological Association. Journal website: <http://www.apa.org/journals/ccp/>

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

Towards improved statistical reporting in the *Journal of Consulting and Clinical Psychology*

Fiona Fidler¹, Geoff Cumming¹, Neil Thomason², Dominique Pannuzzo¹, Julian Smith¹, Penny Fyffe¹, Holly Edmonds¹, Claire Harrington¹ and Rachel Schmitt¹

1 = School of Psychological Science, La Trobe University, Bundoora,
Victoria, 3086, Australia

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

Author contact details:

Fiona Fidler

Department of History and Philosophy of Science

University of Melbourne, Melbourne, Victoria 3010, Australia

email: fidlerfm@unimelb.edu.au

Abstract

In 1997, Philip Kendall's editorial encouraged authors in *JCCP* to report effect sizes and clinical significance. The present authors assessed the influence of that editorial, and other APA initiatives to improve statistical practices, by examining 239 *JCCP* articles published from 1993 to 2001. For ANOVA, reporting of means and standardized effect sizes increased over that period, but the rate of effect size reporting for other types of analyses surveyed remained low. Confidence interval reporting increased little, reaching 17% in 2001. By 2001 the percentage of articles considering clinical (not only statistical) significance was 40%, compared with 36% in 1996. In a follow-up survey of *JCCP* authors ($N=62$), many expressed positive attitudes toward statistical reform, but gave little indication that they understood what was involved. Substantially improving statistical practices may require stricter editorial policies and further guidance for authors on reporting and interpreting measures.

In 1997 Philip Kendall, then *Journal of Consulting and Clinical Psychology (JCCP)* editor, wrote: “Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the required effect size but also a consideration of clinical significance” (Kendall, 1997, p.3). His was one of many calls for statistical reform in psychology, that date back to the 1960s (e.g., Bakan, 1966, 1967; Cohen, 1962; Rosenthal and Gaito, 1963; Rozeboom, 1960).

The view of some reformers is that “There is only one force that can effect a change, and that is the same force that helped institutionalize null hypothesis testing as the *sine qua non* for publication, namely, the editors of major journals” (Sedlmeier & Gigerenzer, 1989, p.315). Kendall’s call for change was thus particularly noteworthy because it was an editorial of a major American Psychological Association (APA) journal. However, some previous journal surveys have shown that editorial policy is not sufficient to substantially change researchers’ practices.

Surveying the Effectiveness of Editorial Policy

In 1993, editor-elect of *Memory & Cognition*, Geoffrey Loftus, aimed “to decrease the overwhelming reliance on hypothesis testing.” (Loftus, 1993, p.3). He encouraged the use of figures with error bars, and the omission of Null Hypothesis Significance Testing (NHST), in manuscripts submitted to *Memory & Cognition*. Finch et al. (2004) found that during Loftus’ term the proportion of articles reporting error bars (whether CIs or SE bars) increased from 7 to 41%. Whilst a substantial achievement, still less than half of authors followed Loftus’ recommendation (the proportion peaked at 41%). Furthermore, NHST was reported in almost every article and it, rather than figures with error bars, was usually used as the basis for interpretation. Loftus’ efforts were, however, those of a lone editor working before any APA interventions (e.g. APA, 1994; Wilkinson et al, 1999; APA, 2001).

However, a number of other journal surveys, including Kirk (1996), Vacha-Haase, Nilsson, Reetz, Lance and Thompson (2000), and Finch, Cumming and Thomason (2001), concurred in finding little influence of the recommendations in the *APA Publication Manual* (APA, 1994). Despite these discouraging results, we might still hope that individual editorial policy, in-conjunction with APA recommendations, might effect a change that, alone, neither has managed to produce.

Studying JCCP

Vacha-Haase et al. (2000) reported that only 5 of 50 editorials published between 1990 and 1998 in 28 APA journals addressed statistical reporting practice. Kendall’s was one.

It is certainly noteworthy that the five editorials we have cited (cf. Kendall, 1997; Murphy, 1997) were even published...these editorials indicate some nascent willingness within one organization to go beyond the ineffective ‘encouragements’ of the 1994 *APA Publication Manual*.” (p.421).

Thus, the *JCCP* is an ideal case to assess the influence of editorial policy on statistical practice. Our study also updates Dar, Serlin and Omer’s (1994) survey of *JCCP* articles published in the 1960s, 1970s and 1980s.

We surveyed *JCCP* articles to assess the response to four pieces of statistical advice (discussed below): (a) statistical guidelines in the 4th edition of the *APA Publication Manual* (1994), (b) Kendall’s 1997 *JCCP* editorial, (c) the APA Task Force on Statistical Inference (TFSI) report (Wilkinson et al., 1999), and (d) a special section on clinical significance published

in *JCCP* in 1999. In addition we emailed *JCCP* authors published during 2000 and 2001, asking about their reactions to Kendall's policy and other statistical reform recommendations.

The APA Publication Manual. The 4th edition of the *APA Publication Manual* (1994) included two new statistical recommendations: It encouraged use of effect sizes and statistical power. (The 5th edition of the *APA Publication Manual* (2001) recommended CIs, noting that they are “in general, the best reporting strategy” (p.22), but our study was completed too early to detect any impact of that recommendation.) Effect sizes were recommended because they help make sense of the importance of individual results. They should also underpin any review or meta-analysis. (*P* values do not provide direct information about the magnitude of effects.) Statistical power was recommended because it gives important a priori information about experimental precision. A statistically non-significant result may well be a result of low power, but without information on precision such results are uninterpretable.

Kendall's editorial. As editor of a major APA journal, Kendall felt a responsibility to inform would-be authors in *JCCP* of advances in the discipline:

During my tenure as editor of *JCCP*, APA has updated its *Publication Manual* and had a committee [the TFSI] look into statistical reporting...As an editor, and consistent with the *Manual* and committee report, I have encouraged authors to report effect sizes and to examine clinical significance. (P. Kendall, personal communication, February 10, 2002).

In this, Kendall was an unusual editor. Most APA editors never informed their authors of these changes (Vacha-Haase et al., 2000), much less tried to encourage their adoption.

The APA Task Force on Statistical Inference. In 1996 the TFSI was appointed to investigate the controversy around NHST. Like Kendall, they recommended that authors discuss clinical significance: “Distinguishing statistical significance from theoretical significance (Kirk, 1996) will help the entire research community publish more substantial results” (Wilkinson et al, 1999, p. 603). They also recommended use of figures, especially with error bars, and strongly endorsed reporting effect sizes: “We must stress again that reporting and interpreting effect sizes...is essential to good research” (p.599); and CIs: “Interval estimates should be given for any effect sizes involving principal outcomes” (p.599).

Crucially, CIs provide information on effect sizes and experimental precision (APA, 2001), yet in psychology they remain relatively little-used (Finch et al., 2001; Kieffer, Reese, & Thompson, 2001). Because researchers rarely see CIs reported, some may not understand why they are important, or they may falsely believe that CIs are equivalent to NHST. CIs can be used to perform NHST, by noting whether the null value is within the CI. But they focus attention on likely effect sizes in a way that *p* values do not. They provide vital information on precision: CI width is a guide to this. Furthermore, CIs “support meta-analysis and meta-analytical thinking focused on estimation” (Cumming & Finch, 2001, p.534).

The JCCP special section on clinical significance. As Chambless and Hollon (1988) pointed out: “If a treatment is to be useful to practitioners it is not enough for treatment effects to be statistically significant: they also need to be large enough to be clinically meaningful” (p.11).

In 1999 the *JCCP* ran a special section on clinical significance. Articles discussed methods, measures and definitions of clinical significance as well as the associated conceptual difficulties and challenges of assessing clinical significance. Some included ‘how to’ guides to calculate various measures (e.g., Gladis, Gosch, Dishuk, & Crits-Christoph, 1999; Jacobson, Roberts, Berns, & McGlinchey, 1999; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). Kazdin (1999) provided an overview.

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

Study 1: Journal Survey

Method

We coded 239 articles that reported new, empirical data (we did not code theoretical articles, meta-analyses or other review studies) from four time periods:

Period 1: 59 articles published in 1993, prior to the release of the 4th edition of the *APA Publication Manual* in July 1994;

Period 2: 59 articles from 1996, submitted after the release of the 4th edition of the *APA Publication Manual* and published prior to the commencement of Kendall's editorship in 1997;

Period 3: 61 articles from 1998 and 1999, submitted during Kendall's editorship and accepted for publication prior to the special section on clinical significance in June 1999; and

Period 4: 60 articles from 2000 and 2001, submitted after publication of the 1999 special section on clinical significance and the TFSI report.

For each article we coded the first occurrence of the following items; we did not record multiple instances.

Type of analysis. Preliminary coding of *JCCP* indicated that ANOVA, chi-square and *t* tests were the most common types of analysis. We therefore limited coding of test statistics, *p* values and effect sizes to these three types of analysis.

Test statistics and p values. For the types of analysis listed above, we recorded whether the appropriate test statistic was reported (e.g., *F* for ANOVA). We also recorded whether *p* values were reported and whether statistically significant results were recorded in the same way as statistically non-significant results (or whether the abbreviation *ns* replaced *p* values).

Effect size. We coded whether or not a standardized or units-free effect size was recorded with an ANOVA (e.g., *d*, η , η^2 , ω , ω^2), chi-square (e.g., Cramer's ϕ , Cochran's *Q*, λ , *W*) or *t* test (e.g., Cohen's *d* or δ). If an ANOVA was present, we coded whether or not means or mean differences were reported for main effects. Often means appeared in large tables, well removed from relevant comments in text. Unclear labeling in some tables also created uncertainty. Because early coding lacked sufficient reliability we coded all articles twice for this item and required agreement between coders for each entry. We limited recording of means to ANOVA only.

Confidence intervals. We recorded any instance of a CI. When present, we noted whether it was reported in a figure, or in a table or text, and whether the confidence level was reported. We also noted what the CI was for (e.g., mean, odds ratio) and, finally, whether it was interpreted (e.g., comments about the width of the interval, or overlap between CIs).

Figures. We recorded any figure displaying data, and whether figures had error bars (e.g., SD, SE, CI).

Clinical significance. To develop criteria for coding clinical significance we searched articles in the 1999 special section on clinical significance to identify key words. Our key word list is presented in Table 1.

It was sometimes difficult to determine whether authors were discussing clinical significance or were conflating clinical and statistical significance. Our coding was lenient in this respect. Because of early problems with reliability we double-coded all articles for this item and required agreement between coders for each entry. Given that the editorial and the special section were more directly related to clinical significance than were the other two interventions, and the time involved in coding this item, we coded clinical significance only for Period 2 (pre Kendall's 1997 editorial) and Period 4 (post 1999 special section).

Table 1.

Key Words for Coding Clinical Significance

clinically/clinical	relevant/relevance
practically/practical	significance/significant
psychologically/psychological	meaningful
	important/importance
	Reliable Change Index
	Return to normal (functioning) / Indistinguishable from normal

Note: Use of any word in the left hand column in conjunction with any word in the right hand column fulfilled the criteria for clinical significance, as did any mention of the phrases in the last two rows.

Coding reliability. We independently cross-coded a randomly selected 20% of the articles. The accuracy of the original coding for all but two items was 90% or more (with discrepancies due to items missed by the original coder, rather than disagreements about coding criteria). Where the accuracy fell below 90% (i.e., the presence or absence of means for ANOVA, and clinical significance), we double-coded all articles for these items. The double coding was first done independently and then, where disagreements arose, in consultation, with a final coding made only after agreement was reached.

Results

Types of analysis. Half of all articles (50%, 120 of 239) reported at least one ANOVA, 53% (126) reported chi-square analysis, and 45% (108) reported at least one *t* test. Most articles used more than one of these methods; 16% (39) did not use any of the three.

Test statistics. The reporting rate of test-statistics was high: 93% (112 of 120) for articles with ANOVA, 94% (118 of 126) for chi-square tests and 90% (97 of 108) for *t* tests.

P values. Virtually all (99%, 197 of 200) of ANOVA, chi-square and *t* test articles reported at least one *p* value. Of the articles reporting these analyses, about a quarter (27%, 54 of 200) used the abbreviation *ns* at least once, rather than reporting *p* values for statistically non-significant results.

Effect sizes. Overall, standardized or units free effect sizes were reported for 32% of articles with ANOVA, 13% with chi-square tests, and 15% with *t* tests. Figure 1 shows the percentage (with 95% CI) for each type of analysis in each period.

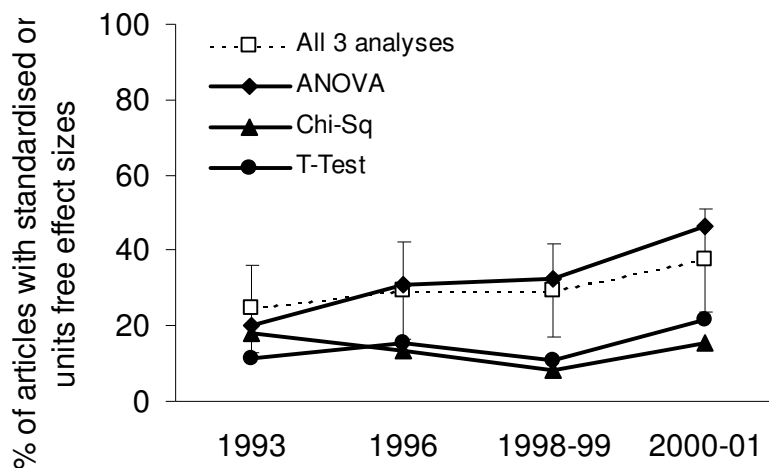


Figure 1.

Experiment 1: Percent of ANOVA, chi-square and *t*-test articles in *JCCP* that reported standardized or units-free effect sizes. Bars are 95% CIs.

The use of these effect sizes with ANOVA rose consistently across the time surveyed, a total increase of 26 percentage points (95% CI = 3 to 47) from 20% in 1993, to 46% in 2001. This increase is statistically significant at the .05 level. Changes in effect size reporting for *t* tests (increase of 10 points, 95% CI = -11 to 30) and chi-square (decrease of 2 points, 95% CI = -24 to 16) were smaller than for ANOVA and did not reach statistical significance. Note however that all three CIs are wide, reflecting a lack of precision in this aspect of the study.

Table 2 shows the percentage of ANOVAs reported with and without at least one mean. Over the first three periods, the percentage of ANOVA articles reporting a mean shows virtually

no change (Period 1 = 60%, Periods 2 and 3 = 58%). By period 4 there were about double the number of articles with ANOVA, and the percentage with a mean increased to 82%. Table 2 also shows reporting rates of ANOVAs missing at least one mean: The percentage is similar over the first three periods (average 63%), then markedly lower in Period 4 (23%).

Table 2.
Percentage of ANOVA Articles Reporting Means for Main Effects

	With at least one mean	Missing at least one mean
1993	69 (24 of 35)	54 (19 of 35)
1996	58 (15 of 26)	65 (17 of 26)
1998-99	58 (18 of 31)	71 (22 of 31)
2000-01	82 (37 of 60)	23 (14 of 60)

Note: Totals do not sum. In many articles there were instances both of main effects reported with means, and of other main effects reported without means.

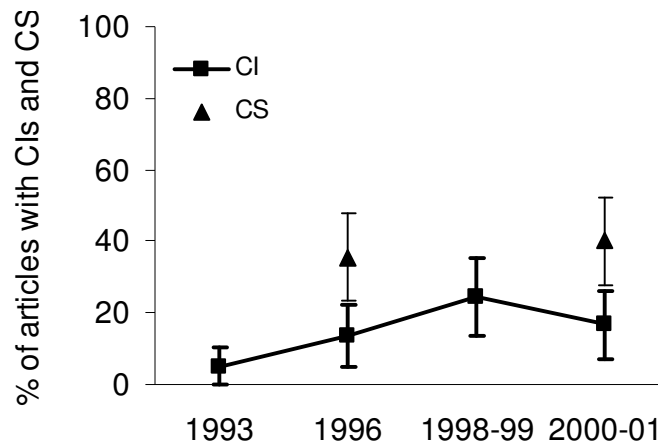


Figure 2.
Experiment 1: Percent of all JCCP articles reporting CIs (Period 1-4), and clinical significance (for Period 2 and 4 only). Bars are 95% CIs.

Confidence intervals. There was some increase in CI reporting over time (see Figure 2), although even in recent articles the reporting rate was low (17%). Most CI reports were in text (72%, 26 of 36), 33% (12) were in tables and in 11% (4) of cases appeared as error bars in figures. (Percentages do not sum to 100% here. Some articles contained more than one type of CI reporting.) CIs were most often reported for odds ratios (15 of 36) and means (11), rarely for correlation coefficients (3) or regression coefficients (1). They were never reported for standardized or other units-free effect size measures. The confidence level of the interval (e.g., 95%, 90%) was not identified in 36% (13 of 36) of articles. Very few authors (11%, 4 of 36)

interpreted the reported CIs. In sum, only about 2% (4 out of 239) of all articles used CIs to interpret their data.

Figures. About a third (31%, 75 of 239) of articles with new empirical data included graphical representations of results. Of these, 5% (4 of 75) included error bars.

Clinical significance. In Period 2, 36% (21 of 59) of all articles considered clinical significance, according to our criteria. In Period 4, after both Kendall's editorial and the special section on clinical significance in *JCCP*, 40% (25 of 60) considered clinical significance. Figure 2 shows these percentages with 95% CIs.

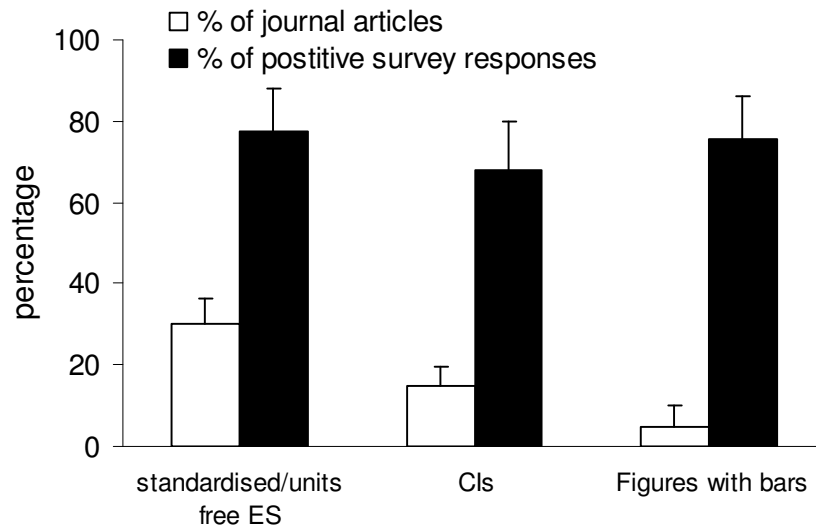


Figure 3.

Experiment 2: Percent of authors' positive responses to reform recommendation survey questions, and percent of *JCCP* articles reporting those same measures, in 2000-01. Bars are upper half 95% CIs. The percentage of Period 4 ANOVA, chi-square and *t* test articles (n=48) reporting standardized and units-free effect sizes is graphed with responses (n=62) to: "Standardized effect sizes (such as Cohen's *d* for *t* tests, Eta-Squared for ANOVA, Cramer's phi for chi-square) are appropriate to my research." The percentage of all Period 4 articles (n=60) reporting CIs appears with responses (n=59) to: "In most cases, it is more useful and informative to report a Confidence Interval instead of a *p* value." The percentage of Period 4 articles with figures (n=17) that included error bars appears with responses (n=60) to: "Graphs that include error bars (i.e., standard error bars or confidence intervals) are preferable to graphs without bars."

Discussion

Our results show some changes that may be responses to calls for statistical reform in *JCCP*. Kendall (1997) asked for "the required effect size" (p.3), which we take to mean the effect size appropriate in the research situation. For ANOVA main effects this will be the means, or mean differences and/or corresponding units-free measures. There was a notable increase

(from 20 to 46%) in the reporting of standardized or units free effect sizes for ANOVA. Further, in 2000-2001 the percentage of ANOVA articles reporting at least one mean was higher than in earlier periods, and the percentage missing at least one mean was lower. These changes are important and promising indicators of reform. However, for chi-square and *t* tests there was very limited change in effect size reporting.

The frequency of reporting of clinical significance was similar for the two periods (36% in Period 2 and 40% in Period 4). However, our coding criteria may have hidden some improvement in the way clinical significance was discussed. Kendall explained (personal communication, April 9, 2001) that prior to his 1997 editorial and the special section, authors would frequently misuse the term 'clinical' to describe no more than 'statistical' significance. Kendall believes this kind of misuse has declined, and that authors are now using more sophisticated measures of clinical significance. Our coding criteria, which relied heavily on word searches, may not always have differentiated between appropriate and inappropriate uses of clinical significance terms. Therefore the percentage of articles in Period 2 genuinely considering clinical significance may be less than the 36% we report here. Dar et al. (1994) reported 30% of articles reporting clinical significance in the 1980s, however, they did not make their coding criteria explicit.

Still, our coding in Period 4 shows that, by our criteria, only 40% of articles made any attempt to discuss clinical significance. This is serious. In a major journal dedicated to research of psychotherapy and other interventions, clinical significance should be relevant to more than 40% of articles.

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

Study 2: Author Survey

Method

We emailed all 2000 and 2001 authors who provided an email address in their article. Forty-seven of the 214 emails sent were returned undelivered, 62 of the remaining 167 authors replied to the survey (although some occasionally skipped a question), a response rate of 37%. Authors were asked a variety of questions about their awareness of and attitude to statistical reform.

Results

Awareness of statistical reform. About half (52%, 31 of 60) of respondents reported that they were “aware” or “vaguely aware” (hereafter aware) of Kendall's 1997 editorial (5 answered “cannot remember” and the remainder were “not aware”). Similarly, 54% (33 of 61) were aware of the TFSI report and 63% (38 of 60) were aware of the 1999 *JCCP* special section on clinical significance. Most (80%, 50 of 62) were aware of at least one of these reform initiatives, 37% (23 of 62) were aware of two, and 26% (16 of 62) of all three.

Effect sizes. Many respondents (77%, 48 of 62) “agreed” or “strongly agreed” (hereafter agreed) with the statement: “Standardized and units free effect sizes (such as Cohen’s *d* for *t*-tests, Eta-squared for ANOVA, Cramer’s Phi for chi-square) are appropriate to my research.”

Almost as many (69%, 42 of 61) agreed that “Standardized effect sizes are easily calculated for the type of data and designs I usually work with.” Similarly, 70% (42 of 60) “disagreed” or “strongly disagreed” (hereafter disagreed) with the statement: “Researchers reporting effect sizes should simply report the appropriate values and leave the reader to make their own interpretation and judgement.” Only 12% (7 of 59) of respondents agreed that “Means, percentages etc are more useful indicators of effect magnitude than standardized measures” (36 disagreed and 16 were “unsure”).

Confidence Intervals. Two-thirds (68%, 40 of 59) of respondents agreed with the statement: “In most cases, it is more useful and informative to report a Confidence Interval instead of a p value”. Fifty-six percent (33 of 59) agreed that “Confidence Intervals are easily calculated for the type of data and designs I usually work with” and 57% (35 of 61) agreed that “Reporting Confidence Intervals is relevant in the sort of research I usually do.”

Figures. Fifty-eight percent (35 of 60) agreed with the statement: “Results reported in figures or graphs are, in most cases, easier to read and interpret than tables”. Three-quarters (75%, 46 of 61) agreed that “Graphs that include error bars (i.e. standard error bars or confidence intervals) are preferable to graphs without bars.” Fully 80% (49 of 61) disagreed with the statement: “Graphs that include error bars are complicated and difficult to interpret.”

Clinical significance. We asked participants: “In your opinion, to what extent do authors in your field consider in their papers the clinical significance of their results?” Just one respondent replied that it was “always” considered. One quarter (25%, 15 of 60) judged that clinical significance was “often” considered, 47% (28 of 60) thought it was “sometimes” considered and 27% reported it was “rarely” considered. No participants replied that it was “never” considered.

Attitudes and reporting. Respondent anonymity prevented us from matching survey responses to articles published by an individual. However, comparisons between general attitudes and general reporting practices are possible. Figure 3 shows percentages of positive responses to survey questions alongside the Period 4 reporting rate of corresponding measures. In each case, we chose the survey question that most closely matched the reporting measure. Respondents had a generally accurate appreciation that many articles did not discuss clinical significance, but on other issues there were substantial discrepancies between attitudes and reporting practices. Respondents’ attitudes towards effect sizes are much more positive than reporting practices would predict; so are attitudes towards CIs. The discrepancy between attitude and practice was most pronounced in relation to figures with bars.

Discussion

A large majority (80%) of authors were aware of at least one of these three reform initiatives: Kendall’s editorial, the TFSI report and the *JCCP* special section. Attitudes towards standardized and units-free effect sizes were positive (77% thought they were appropriate to their research), as were attitudes towards CIs (68% thought they were more useful and informative than p values). This is seemingly good news for reform, although we acknowledge that respondents may have been more sympathetic to statistical reform than non-respondents.

Further, despite our efforts to chose neutrally-worded questions, respondents may have felt social desirability pressure to give reform-positive responses.

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

General Discussion

By informing potential authors of desirable statistical practices Kendall took, as editor of *JCCP*, an important and unusual step. Our results suggest, however, that his policy has been, at best, only partly effective in changing the ways that authors report and interpret their results. Since the rates reported for the 1980s by Dar et al. (1994) there have been some notable improvements: For standardized and units-free effect sizes, from none to 40% (overall for 2000-01 articles in our survey); and for CIs, from none to 17% (Period 4 in our survey). The reporting of clinical significance in *JCCP* has increased but remains alarmingly low: 30% in the 1980s (Dar et al, 1994) and 40% in 2000-01. Perhaps some researchers have been slow to pick up on clinical significance because they feel their results are diminished in such a presentation: It is usually much easier to find a statistically significant result than one that is clinically significant.

Previous surveys (e.g., Finch et al., 2001; Vacha-Haase et al., 2000) had prepared us for the 1994 APA *Publication Manual* having little or no effect on statistical reporting. But what of Kendall's policy? How could such "nascent willingness" (Vacha-Haase, et al., 2000, p. 421) to go beyond the APA *Publication Manual* not pay off? We suggest two possible explanations. First, on effect sizes, the policy was vague. What does "required effect size" mean? Authors may have felt they were meeting this criterion simply by publishing means for each level of an independent variable, rather than the mean difference between levels, or a standardized or units-free effect size, as would be appropriate in many cases. Effect sizes should be salient and accessible, and relate directly to the experimental questions of most interest (Kline, 2004, section II).

A second possible explanation is that the policy was not strongly enforced. When Rothman was editor of *Epidemiology* he wrote the following policy on *p* values: "In *Epidemiology*, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in data analysis" (1998, p.334). In 2000, *Epidemiology* did not publish a single *p* value, and over 90% of articles reported CIs (Fidler et al., 2004).

This is obviously a more radical position than Kendall's, but the nature of the policy is not as relevant here as the fact that enforcement by Rothman and his co-editors was highly effective at changing practices. Kendall's strategy on the other hand was to inform rather than reject: "A paper would not be rejected because of the absence of effect size data, but authors

would be told of the need to report effect sizes” (P. Kendall, personal communication, April 9, 2001).

Recommendations

Enforced policy. Our results suggest that editors need to go further. Thompson has argued the same of the *APA Publication Manual*: “To present an ‘encouragement’ [to report effect sizes] in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message ‘these myriad requirements count, this encouragement doesn’t.’” (1999, p.162) The challenge for editors is to be insistent on improved practices, while allowing scope for new statistical techniques to make their contributions. Of course, this should not be the task of individual editors, like Kendall, alone. The *APA Publication Manual* needs to support these reform efforts, in a more sustained and consistent way than either the 4th or 5th editions do (Fidler, 2002).

Improved education. Our results also demonstrate the need for improved guidelines and education of psychologists and researchers. Results from our author survey demonstrate positive attitudes towards statistical reform recommendations, at least among our respondents. But they also indicate that researchers may lack some relevant knowledge. Perhaps this is not surprising given the conservative nature of many graduate programs in quantitative psychology, and the over-emphasis placed on hypothesis testing (Aiken, West, Sechrest, & Reno, 1990). In addition, recent evidence that many researchers hold misconceptions about CIs (Belia, Fidler, Williams, & Cumming, 2005; Cumming, Williams, & Fidler, 2004) further emphasizes the need for better education and guidelines in relation to reform techniques.

Reform resources

We refer readers to the following sources as guides to using and interpreting statistical reform recommendations. The TFSI report (Wilkinson et al., 1999) is a good starting point and includes many references. Kline (2004) gives excellent practical guidance on a wide range of issues. These sources also make clear that there is much more to reformed statistical practice than the use of effect sizes, CIs, and clinical significance, which have been our focus in this article.

Kirk (1996) described 40 measures of effect magnitude. Rosnow and Rosenthal (2003) also provided a useful summary of many effect sizes. Rosenthal (1994) and Rosenthal and Rubin (e.g., 1982, 1986, 1994) have written extensively about effect sizes. Becker (2003) introduced a set of articles that gave a more advanced discussion.

Cumming and Finch (2001) provided a primer on the calculation and use of CIs. CIs for some effect sizes, including Cohen’s d , η^2 and R^2 , require non-central distributions. Methods for calculating these CIs are explained in Steiger and Foudali (1997), Cumming and Finch (2001), Fidler and Thompson (2001), Steiger (2004) and Smithson (2001, 2002). Guidance and software for calculation of CIs for many measures and designs (including odds ratios from logistic regression) is provided by Altman, Machin, Bryant and Gardner (2000). Cumming and Finch (2001, 2005) discuss how CIs can be represented, and used to support interpretation of results.

The *JCCP* special section (1999) is an excellent starting point for advice on clinical significance. Of course, recommendations to differentiate clinical or substantive significance from statistical significance predate Kendall’s policy (e.g., Grove & Meehl, 1996; Kendall &

Grove, 1988; Jacobson, Follette & Revenstorf, 1984; Jacobson & Truax, 1991; Lees & Neufield, 1994; Meehl, 1954; Rosenthal, 1983). In 1988 a special issue of *Behavioral Assessment* was devoted to defining clinically significant change. There has also been widespread discussion of clinical significance in the medical literature (e.g., Daly, 2000; Lindgren, Wielinski, & Finkelstein, 1994; Luus, Muller, & Meyer, 1989; Manchanda, 1986). More recent articles in psychology include Ogles, Lunnen, and Bonesteel (2001), and Beutler and Moleiro (2001).

Limitations of this study

One limitation of our survey of *JCCP* is that it was restricted to ANOVA, chi-squared and *t*-tests. As discussed, these were the three most popular techniques used in the journal, however, a more complete analysis would include regression and other methods. One could also argue that the measures of improved statistical reporting we have chosen are inadequate. We make no claim that using effect sizes, CIs and reporting clinical significance are the only strategies for improving statistical reporting, beyond NHST. There are a variety of modeling techniques, likelihood analysis, Bayesian methods and others techniques that should be considered. Our focus, however, was on the primary recommendations of the TFSI and Kendall's policy.

Conclusion

The 5th edition APA *Publication Manual* may be more successful than individual editorial policies or the TFSI report have been. However, as we have previously noted (Fidler, 2002; Finch et al., 2002), in our opinion it did not go nearly far enough. It recommended effect sizes and CIs, but failed to provide examples of how to report and interpret them, even though it provided several examples of how to report *p* values. We hope that the 6th edition, when released, will be more helpful to those who want to report the statistics that the APA *Publication Manual* itself recommends. Statistical reform is important, and has the potential to improve markedly the effectiveness of research, and research communication. Achieving statistical reform will require a range of strategies, including commitment by researchers, better guidance for authors, stronger requirements by the APA *Publication Manual*, and insistence by journal editors that recommendations be followed.

References

- Aiken, L., West, S., Schrest, L., & Reno, R.R. (with Roediger, H. L., Scarr, S., Kazdin, A. E. & Sherman, S. J.). (1990). The training in statistics, methodology, and measurement in psychology. *American Psychologist*, *45*, 721-735.
- Altman, D.G., Machin, D., Bryant, T.N., & Gardner, M.J. (Eds). (2000). *Statistics with Confidence* (2nd ed.). London: BMJ Books.
- 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. (1994). *Publication Manual of the American Psychological Association* (4th ed.). Washington, DC: Author.
- American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437.

- Bakan, D. (1967). *On Method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass.
- Becker, B. J. (2003). Introduction to the special section on metric in meta-analysis. *Psychological Methods*, 8, 403-405.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*.
- Beutler, L. E., & Moleiro, C. (2001). Clinical versus reliable and significant change. *Clinical Psychology: Science & Practice*, 8, 441-445.
- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.
- Chambless, D.L., & Hollon, S.D. (1988). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, 66, 7-19.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.
- 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.
- Daly, L. (2000). Confidence intervals and sample sizes. In D. G. Altman, D. Machin, T. N. Bryant, & M. J. Gardner (Eds.) *Statistics with confidence* (2nd ed. pp.139-152). London: BMJ Books.
- Dar, R., Serlin, R. C., & H. Omer. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology*, 62, 75-82.
- Fidler, F. (2002). The fifth edition of the *APA Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, 62, 749-770.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals but they can't make them think: Statistical reform lessons from medicine. *Psychological Science*, 15, 119-126.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, 61, 575-604.
- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, 61, 181-210.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future APA guidelines for statistical practice. *Theory and Psychology*, 12, 825-853.

- 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.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293-323.
- Gladis, M. M., Gosch, E.A., Dishuk, N. M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 320-331.
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7.
- Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336-352.
- Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300-307.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12-19.
- Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 332-339.
- Kendall, P.C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *65*, 3-5.
- Kendall, P.C., & Grove, W. (1988) Normative comparisons in therapy outcome. *Behavioral Assessment*, *10*, 147-158.
- Kendall, P. C., Marrs Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 285-299.
- Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, *69*, 280-309.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213-218.
- Kline, R. B. (2004) *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.
- Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*, 635-643.
- Lees, M. C., & Neufeld, R. W. J. (1994). Matching the limits of clinical inference to the limits of quantitative methods: A formal appeal to practice what we consistently preach. *Canadian Psychology*, *35*, 268-282.
- Lindgren, B. R., Wielinski, C. L., & Finkelstein, S. M. (1994). Contrasting clinical and statistical significance within the research setting. *Pediatric Pulmonology*, *18*, 64.

- Loftus, G. R. (1993). Editorial comment. *Memory & Cognition*, 21, 1-3.
- Luus, H. G., Muller, F. O., & Meyer, B. H. (1989). Statistical significance versus clinical relevance. Part II. The use and interpretation of confidence intervals. *South African Medical Journal*, 76, 626-629.
- Manchanda, R. (1986). Criteria for measuring change: Statistical significance vs clinical significance. *British Journal of Psychiatry*, 148, 744-745.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Northvale, NJ: Jason Aronson.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.
- Oakes, M. W. (1986). *Statistical inference: a commentary for the social and behavioural sciences*. Chichester, U.K.: Wiley.
- Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review*, 21, 421-446.
- Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology*, 51, 4-13.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.) *Handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
- Rosenthal, R. & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 55, 33-38.
- Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166-169.
- Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, 99, 400-406.
- Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, 5, 329-334.
- Rosnow, R., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, 57, 221-237.
- Rothman, K. (1998). Writing for *Epidemiology*. *Epidemiology*, 9, 333-337
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Schmitt, N. (1989). Editorial. *Journal of Applied Psychology*, 74, 843-845.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-315.
- Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164-182.
- Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 222-257). Hillsdale, NJ: Erlbaum.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Methods*, 61, 605-632.

- Smithson, M. (2002). *Confidence Intervals*. Thousand Oaks, CA: Sage.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.
- Thompson, B. (1999). If statistical tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology*, 9, 165-181.
- 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 thank Sue Finch for valuable suggestions and ideas, and Matthew Klugman for comments on a draft of this manuscript, and acknowledge the support of the Australian Research Council. They may be contacted by email: Fiona Fidler: fidlerfm@unimelb.edu.au; Geoff Cumming: g.cumming@latrobe.edu.au.