

Running head: VALUE OF RCT EVIDENCE DEPENDS ON STATISTICAL ANALYSIS

Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research and Therapy*, 46, 270-281.

© Elsevier B. V. Journal website:

[http://www.elsevier.com/wps/find/journaldescription.cws\\_home/265/description#description](http://www.elsevier.com/wps/find/journaldescription.cws_home/265/description#description)

This article may not exactly replicate the final version published in the journal. (It is the version just before copy editing.) It is not the copy of record.

### **The value of RCT Evidence Depends on the Quality of Statistical Analysis**

Cathy Faulkner, Fiona Fidler and Geoff Cumming<sup>a</sup>

School of Psychological Science, La Trobe University, Melbourne, Australia  
[s.faulkner@latrobe.edu.au](mailto:s.faulkner@latrobe.edu.au); [f.fidler@latrobe.edu.au](mailto:f.fidler@latrobe.edu.au), [g.cumming@latrobe.edu.au](mailto:g.cumming@latrobe.edu.au)

The authors examined statistical practices in 193 randomized controlled trials (RCTs) of psychological therapies published in prominent psychology and psychiatry journals during 1999 to 2003. Statistical significance tests were used in 99% of RCTs, 84% discussed clinical significance, but only 46% considered—even minimally—statistical power, 31% interpreted effect size, and only 2% interpreted confidence intervals. In a second study, 42 respondents to an email survey of the authors of RCTs analyzed in the first study indicated they consider it very important to know the magnitude and clinical importance of the effect, in addition to whether a treatment effect exists. The present authors conclude that published RCTs focus on statistical significance tests (“Is there an effect or difference?”), and neglect other important questions: “How large is the effect?” and “Is the effect clinically important?” They advocate improved statistical reporting of RCTs especially by reporting and interpreting clinical significance, effect sizes and confidence intervals.

*Key words:* randomized controlled trial; statistical significance; CIs; clinical significance; effect size

Randomized controlled trials (RCTs) are the foundation for the evidence-based movement in clinical psychology. However the information RCTs provide may be limited by the types of statistical analyses reported in published articles. Consolidated Standards of Reporting Trials (CONSORT, n.d.) and the *Publication Manual* of the American Psychological Association (APA, 2001) recommend that results focus on effect size, clinical or practical significance, and interval estimates such as 95% confidence intervals (CIs). (APA; CONSORT; Wilkinson & Task Force on Statistical Inference, 1999).

Most RCT statistics fall into one of three categories: statistics about whether a genuine (population) effect exists (statistical significance tests combined with power, or CIs); statistics indicating the magnitude of an effect (means, standardised or units free effect size, other estimates) and the precision of those estimates (again, CIs, SEs); and statistics examining the

clinical importance of an effect (measures of clinical significance). These categories correspond to three key research questions (Kirk, 2001, suggested three similar questions):

- Is there a true (population) effect?
- How large is the effect?
- To what extent is the effect clinically (or practically) important?

To the extent that psychology relies on statistical significance tests for interpreting outcomes, only the first question is considered. The discipline is thus reduced to observing ordinal relationships (e.g., therapy > waiting list) rather than estimating effects (e.g., .1 specific therapy effects + .4 general therapy effects + .2 client readiness to change + .1 therapist effectiveness = .8 overall outcome effect size). There is evidence that relying only on statistical significance testing has stunted investigations and distorted important research goals in clinical psychology, and unnecessarily delayed researchers from reaching conclusions in many areas of psychological research (Hunter & Schmidt, 2004; Robbins, 1988; Rossi, 1997). CIs are often recommended as a supplement or replacement of statistical significance tests (e.g., Harlow, 1997), for the following reasons.

First, CIs make uncertainty explicit. By this we mean that CIs offer immediate information about *precision*. A wide interval indicates a lack of precision; a narrower interval, relatively better precision. This means that studies with poor precision cannot be mistaken as evidence for nil effects, one of the major problems associated with *p* values. Second, CIs by definition contain point estimates of effect size (that is what they are constructed around!). When CIs are used to report results, effect size cannot be overlooked. By contrast, when reporting *p* values it is common for researchers to neglect effect size reporting. Third, CIs do not preclude decisions: They can be also be used to reject or fail to reject the null, when appropriate, by noting whether or not the null is captured.

There are perhaps further cognitive advantages of CIs as well. Since CIs rely on the same sample information as significance tests, and belong to the same, frequentist, philosophy of statistics, some may be tempted to think they are ‘the same thing’ as significance tests. Yet, they are different in important ways. The belief that they are the same ignores the extra information about precision that a CI provides, and also dismisses a mass of evidence that different formats of equivalent information can profoundly affect our ability to complete conceptual algorithms and reason using the information (e.g., Gigerenzer & Hoffrage, 1995). Finally, CIs may facilitate meta-analytic thinking (Cumming & Finch, 2001). That is, they may assist thinking across the results of independent studies, acknowledging prior information with an emphasis on effect size, rather than making dichotomous ‘reject’ or ‘fail-to-reject’ decisions based on the outcome of single experiments.

Unfortunately surveys of articles in psychology journals indicate pervasive use of statistical significance tests and neglect of other statistics, despite the recommendations from APA and CONSORT. Table 1 shows the statistical reporting practices used in a broad range of psychology journals, based on weighted averages from 10 journal surveys. All known surveys of psychology journals from the years 1990 to 2006 were included. The weighted averages in Table 1 show that statistical significance tests are used very frequently (88% of articles), but far fewer articles include CIs (11%), standardized or units-free effect sizes (30%), clinical significance (24%) or a statistical power analysis (3%). Even when more informative statistics are reported,

they are rarely used in interpretations. For example, units-free effect sizes are reported in 30% of articles but only interpreted in 8%.

In the current study, we were interested specifically in RCTs, because these are most central for evidence-based practice by clinical psychologists. In the first stage of this study, we surveyed journal articles reporting RCTs of psychological therapies for psychological disorders. In the second stage, we asked authors of RCTs their opinions about the three generic research questions listed above.

Table 1.

*Summary Results From Studies of Statistical Practices in Psychological Journals\**

<i>Reporting Practice Journal Period (Reference)</i>	<i>Author(s)</i>	<i>N</i>	<i>% (n) of Articles with the Practice</i>
<i>Report Confidence Intervals</i>			
<i>JCCP</i> 1993-2001	Fidler, Cumming et al (2005)	239	18% (43)
<i>M&amp;C</i> 1998-2000	Finch, Cumming et al (2004)	228	10% (23)
<i>JAP</i> 1999	Finch, Cumming & Thomason (2001)	150	3% (5)
<i>IJED</i> 1990, 2000	Crosby, Wonderlich et al (2006)	152	9% (13)
<i>Overall</i>		769	11% (84)
<i>Interpret Confidence Intervals</i>			
<i>JCCP</i> 1993-2001	Fidler, Cumming et al (2005)	239	2% (5)
<i>Report Effect Size</i>			
<i>JCEN, JINS, N</i> 1998-1999	Bezeau & Graves (2001)	68	9% (6)
<i>JAP, JEdP, JexP, L&amp;M, JPSP</i> 1995	Kirk (1996)	389	42% (163)
<i>JCD</i> 1996	Thompson & Snyder (1998)	25	60% (15)
<i>PPRP</i> 1990-7	Vacha-Haase & Ness (1999)	265	20% (53)
<i>MECD</i> 1990-6	Vacha-Haase & Nilsson (1998)	83	35% (29)
<i>IJED</i> 1990, 2000	Crosby, Wonderlich et al (2006)	152	17% (27)
<i>JCCP</i> 1993-2001	Fidler, Cumming et al (2005)	200	30% (60)
<i>Overall</i>		1182	30% (353)
<i>Interpret Effect Size</i>			
<i>JCD</i> 1996	Thompson & Snyder (1998)	25	8% (2)

<i>Discuss Clinical Significance</i>			
<i>JCCP</i> 1993-2001	Fidler, Cumming et al (2005)	239	36% (86)
<i>IJED</i> 1990, 2000	Crosby, Wonderlich et al (2006)	152	6% (9)
<i>Overall</i>		391	24% (95)
<hr/>			
<i>Use Statistical Significance Tests</i>			
<i>JAP</i> 1990-4	Sedlmeier & Gigerenzer (1989)	80	94% (75)
<i>MECD</i> 1990-6	Vacha-Haase & Nilsson (1998)	83	82% (68)
<i>PPRP</i> 1990-7	Vacha-Haase & Ness (1999)	265	77% (204)
<i>IJED</i> 1990, 2000	Crosby, Wonderlich et al (2006)	152	95% (144)
<i>M&amp;C</i> 1998-2000	Finch, Cumming et al (2004)	228	97% (221)
<i>Overall</i>		808	88% (712)
<hr/>			
<i>Report a Power Analysis</i>			
<i>JCEN, JINS, N</i> 1998-1999	Bezeau & Graves (2001)	68	3% (2)
<i>JAP</i> 1999	Finch, Cumming & Thomason (2001)	30	10% (3)
<i>IJED</i> 1990, 2000	Crosby, Wonderlich et al (2006)	152	2% (3)
<i>Overall</i>		250	3% (8)

*IJED = International Journal of Eating Disorders; JAP = Journal of Applied Psychology; JCCP = Journal of Consulting and Clinical Psychology; JCD = Journal of Counseling and Development; JCEN = Journal of Clinical and Experimental Neuropsychology; JEdP = Journal of Educational Psychology; JExP = Journal of Experimental Psychology; JINS = Journal of the International Neuropsychological Society; JPSP = Journal of Personality and Social Psychology; L&M = Learning and Memory; M&C = Memory and Cognition; MECD = Measurement and Evaluation in Counseling and Development; N = Neuropsychology; PPRP = Professional Psychology: Research and Practice.*

### Study 1: Analysis of Published RCTs

We surveyed the statistical reporting practices of RCTs of psychological therapies for psychological disorders. RCTs were sought from clinical psychology journals and psychiatry journals. Psychiatry journals were included for comparison because general medicine has adopted CIs to a greater extent than psychology (Fidler, Cumming, Burgman, & Thomason, 2004). We also investigated the research designs used in the published RCTs, so that future recommendations on statistical practices can be tailored to typical designs common in RCTs.

### Method

We selected journals of importance to clinical psychology and psychiatry using the following criteria: (a) Journals had an impact factor of two or more, from the *Journal Citation Reports (JCR)*, Social Sciences Edition. (b) Journal content was specific to clinical psychology or psychiatry. This was determined from journals' stated aims and scope. (c) Journal content was otherwise broad, not specific to one age group or disorder (e.g., geriatric psychiatry, schizophrenia). (d) Journals contained RCTs of psychological therapies.

Of the 161 journals considered, two clinical psychology journals met criteria for inclusion: *Journal of Consulting and Clinical Psychology*, and *Behaviour, Research and Therapy*. Six psychiatry journals met criteria: *Archives of General Psychiatry*, *American Journal of Psychiatry*, *British Journal of Psychiatry*, *Journal of Clinical Psychiatry*, *Psychological Medicine*, and *Acta Psychiatrica Scandinavica*.

We examined only RCTs of psychological therapies for psychological disorders. Pharmacological studies were only included when they had a comparison with a psychological therapy condition. Overall, 193 RCTs were included, from the years 1999 to 2003, as shown in Table 2. Table 3 shows the variables examined. A detailed coding manual was developed.

Table 2.

*Number of Published RCTs (1999 to 2003) in Journals Analyzed in Study 1*

Clinical Psychology Journals	RCTs
<i>Journal of Consulting &amp; Clinical Psychology</i>	77
<i>Behaviour, Research &amp; Therapy</i>	27
Total	104
Psychiatry Journals	
<i>Archives of General Psychiatry</i>	28
<i>British Journal of Psychiatry</i>	26
<i>American Journal of Psychiatry</i>	13
<i>Acta Psychiatrica Scandinavica</i>	9
<i>Psychological Medicine</i>	7
<i>Journal of Clinical Psychiatry</i>	6
Total	89

### *Reliability of Coding*

RCTs were selected and coded by the first author, then a doctoral candidate in clinical psychology. For each journal, 10% of RCTs in each year were also coded independently by the second author. Cohen's Kappa statistics for inter-rater reliability are reported in Table 3, with items coded. Agreement was above 90% for the number of independent conditions, reporting of effect size, CIs and clinical significance; and was at least 80% for all variables but three: The number of measures, number of measurement times, and number of months follow-up. These three variables proved difficult to code reliably, because of complex designs and vague and incomplete information in method sections—perhaps partly reflecting word limits in published papers. For example, we frequently observed sub-scales that were poorly defined in the method but were used in results, measures described outside of the section titled 'measures', and descriptions of the frequency of measurement for various measures that added to an unmentioned

different total of measurement times. There was a wide variation in how follow-up was defined, and it was often described vaguely, requiring us to piece together brief comments made in the method and results. On average, the primary coder was within 2.3 measures, 2.2 measurement times, and 2.0 months follow-up, of the secondary coder.

Table 3.

*RCT Characteristics Examined in Study 1**Characteristics of the Research Design*

\*Types of treatments (e.g., psychosocial therapy, medication, waiting list)

\*Number of participant groups

\*Number of psychosocial therapy types

\*Number of participants randomized

Number of measures (including subscales)

Number of measurement times (e.g., pre-treatment, post-treatment, follow-ups)

Months follow-up after treatment end

Cohen's Kappa (K)=n/a. Percent agreement between coders > 80% where indicated \*.

*Reporting of Results in RCTs*

Reporting and interpreting *confidence intervals* and *standard errors* (in text, table or figures)

$K_{CI \text{ reported (text, fig, table)}}=1$ ;  $K_{SE \text{ reported}}=.97$

Reporting of standardized (e.g., Cohen's *d*) or units-free (e.g., Pearson's *r*, odds ratios) *effect size* (in text, table or figures)

$K_{ES \text{ reported}}=.87$ ;  $K_{ES \text{ type}}=.88$

Interpretation of effect size (e.g., mention of the effect size being small, that the effect size was larger than another, or greater than zero)

$K_{ES \text{ interpreted}}=1$

Discussion of *clinical significance* (i.e., discussing clinically important criterion such as % reduction in symptoms, scoring below a cut-off on a measure, abstinence from an addictive substance, avoiding hospital re-admission, no longer meeting criteria for diagnosis. We also recorded the use of the reliable change index, normative comparisons, or the term 'clinical significance'.)

$K_{CS}=.96$

Use of *figures* (type and number of figures, use of a panel of figures)

$K_{Fig \text{ reported}}=$ ;  $K_{Fig \text{ number}}=n/a$

Reporting of *statistical power* (We coded any recognition of power, from reporting a calculated power estimate, to merely mentioning that sample size was limited.)

$K_{Power \text{ calc}}=.78$ ;  $K_{Power \text{ implied}}=.86$

Reporting of *statistical significance tests* (We coded the presence or absence of significance tests, and whether exact *p* values were used throughout, or whether inexact *p* values were used at least once. We defined exact *p* values as being "*p* = ..." (e.g., *p* = .13, or *p* = .01), non-conventional *p* values not due to Bonferroni adjustments (e.g., *p* < .02), or *p* values stated to three decimal places (e.g., *p* < .001). Inexact *p* values were any use that did not meet the above criteria, for example, *p* < .05, *p* < .01, *ns*, \*\*.)

$K_{NHST \text{ present/absent}}=1$

$K_{NHST \text{ reported}}=.89$

$K_{NHST \text{ exact/inexact}}=.84$

### ***Results Part 1: Design Characteristics of RCTs***

RCTs in clinical psychology journals focused on anxiety disorders (44% of RCTs), and substance use disorders (27%). In contrast, RCTs in psychiatry journals had a greater emphasis on mood disorders (32%) and psychotic disorders (18%).

Frequently used treatment approaches were cognitive behavioral therapy (used in 27% of all RCTs), exposure (16%), and cognitive therapy (11%). The median number of participants was 72 in clinical psychology RCTs ( $M = 86$ ; 95% CI: [75-97] after removing two outliers) and 96 in psychiatry RCTs ( $M = 96$  [84-109] after removing three high outliers).

The most frequent design was two psychological treatment conditions (27% of clinical psychology RCTs, 21% of psychiatry RCTs). The median number of participants per independent condition was 27 in clinical psychology RCTs, and 30 in psychiatry RCTs. RCTs frequently had four measurement times: pre-, during-, post-treatment and follow-up (range: 2 to 115 measurement times). Clinical psychology articles had a median of 11 dependent variables; psychiatry articles had 8. The median follow-up after treatment ended was seven months for clinical psychology RCTs, and six months for psychiatry RCTs. There was no follow-up in 43% of psychiatry RCTs and 13% of clinical psychology RCTs.

### ***Results Part 2: Inferential Statistics Used in RCTs***

Figure 1a shows the rates of reporting and interpreting CIs or standard errors in clinical psychology or psychiatry RCTs. In psychiatry RCTs, CIs or standard errors were reported far more often (difference = 30% [17, 42]), but still rarely interpreted. (CIs for proportions reported here were calculated using the method recommended by Newcombe & Altman, 2000.) Few RCTs included CIs or standard errors in figures (3% [1, 8] in clinical psychology, 7% [3, 14] in psychiatry).

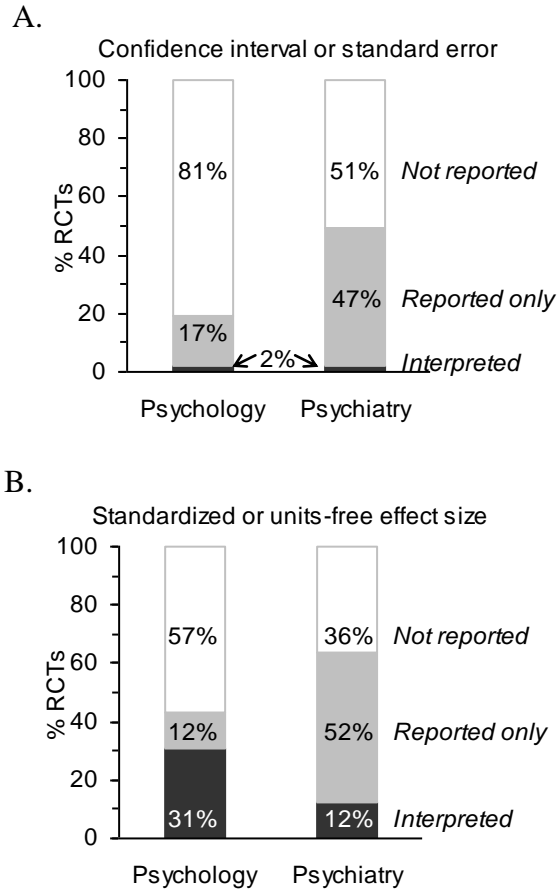
Standardised or units-free effect size were more often reported in RCTs in psychiatry journals (64%) than in RCTs in clinical psychology journals (43%), as shown in Figure 1b. In clinical psychology journals, 31% of those RCTs reporting effect sizes also interpreted them, compared with only 12% interpretations in psychiatry journals.

Most RCTs—84% [75, 90] in each discipline—discussed clinical significance, as shown in Figure 2; 26% [19, 35] of RCTs in clinical psychology journals described criteria for clinical significance and also used multiple approaches to evaluate clinical significance, while only 1% [0, 6] of psychiatry RCTs did so (difference = 25% [16, 34]). This may indicate a particularly thorough approach to exploring clinically significant change in clinical psychology, or alternatively, the existence of more established protocols in psychiatry.

The importance of figures in communicating scientific information has often been discussed (e.g., L. D. Smith, Best, Stubbs, Archibald, & Robertson-Nay, 2002). Yet, figures were used sparingly in RCTs: Less than half (48% [39, 58]) of RCTs in clinical psychology journals contained any figures, and 60% [49, 69] of RCTs in psychiatry journals. Groups of figures, in which two or more figures of similar format were displayed together to facilitate comparisons, were used in 19% [13, 28] of clinical psychology RCTs, and 12% [7, 21] of psychiatry RCTs. Error bars were included in only 10% [6, 15] of figures (15 of 156 figures) in clinical psychology RCTs and 22% [16, 29] of figures (29 of 134 figures) in psychiatry RCTs (difference = 12% [4, 21]). Error bars were standard errors (10% [6, 16] of all figures), uncertain type (no explanation anywhere in the text or figure, 6% [4, 9] of all figures), and CIs (1% [0, 3] of all figures).

Of the 193 RCTs, 99% [96, 100] used at least one instance of statistical significance tests to report results. Most RCTs used inexact  $p$  values (e.g.,  $p < .05$ ,  $p < .01$ ,  $ns$ , \*\*) at least once,

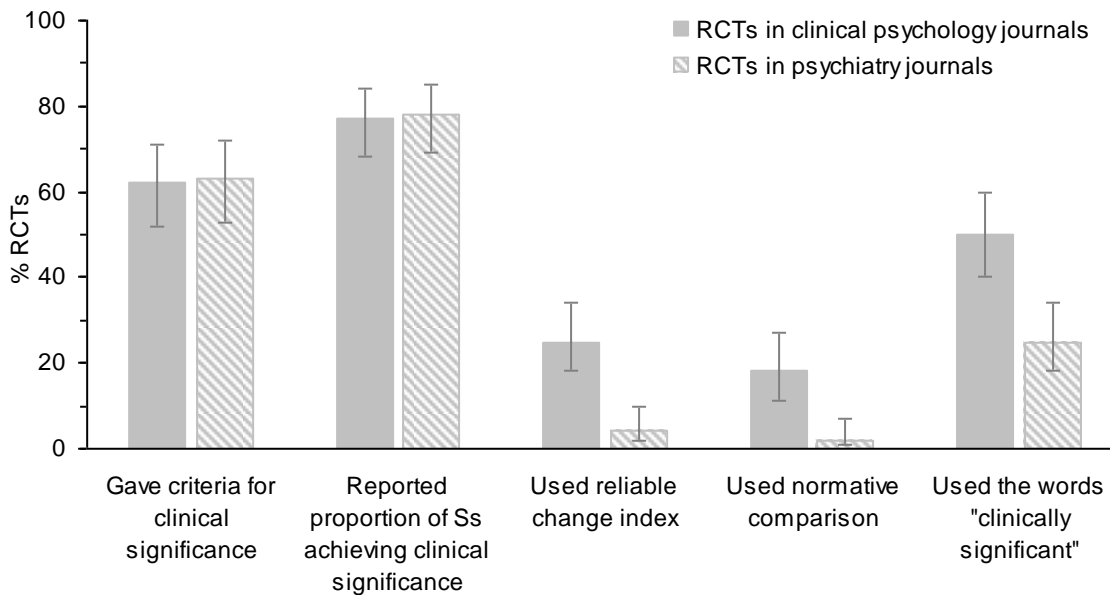
rather than exact  $p$  values throughout (see Figure 3a). We defined exact  $p$  values as being “ $p = \dots$ ” (e.g.,  $p = .13$ , or  $p = .01$ ), non-conventional  $p$  values not due to Bonferroni adjustments (e.g.,  $p < .02$ ), or  $p$  values stated to three decimal places (e.g.,  $p < .001$ ).



*Figure 1.* Study 1. A. Percentage of RCTs reporting and interpreting CIs and standard errors in clinical psychology journals ( $n = 104$  RCTs) and psychiatry journals ( $n = 89$  RCTs). B. Percentage of RCTs reporting and interpreting a standardized or units-free effect size in clinical psychology journals and psychiatry journals.

Statistical power calculations were provided for 10% of clinical psychology RCTs and 28% for psychiatry RCTs. As shown in Figure 3b, approximately half of all RCTs did not mention power, even implicitly. Our coding criterion for implied power was very broad, including any mention that sample size might have impacted on a study’s ability to detect differences using significance tests. We used the reported sample sizes and other published information to calculate that, of the 69 clinical psychology RCTs and 41 psychiatry RCTs that compared two or more psychological therapies, *none* had an a priori power equal or greater than .80 for  $\alpha = .05$  (two-tailed) to detect an effect size of Cohen’s  $d = 0.2$ , which has been well established as the average difference between two psychological therapies<sup>1</sup> (Luborsky et al., 2002; Shapiro & Shapiro, 1982; M. L. Smith & Glass, 1977; Wampold et al., 1997). Further, the

average statistical power for the 110 RCTs was less than .12, and 84% ([76, 89] 92 of 110 RCTs) had statistical power less than .20, to detect a difference of  $d = 0.2$ . Most of these 110 RCTs (75% [66, 82]) did not interpret effect size, and most (56% [46, 64]) did not mention the impact of sample size or power. In other words, many RCTs aimed to investigate the differential effects of two therapies, but most would have found statistically non-significant differences due to low power (not null effects), and yet those RCTs relied on statistical significance to guide interpretation of outcomes.



*Figure 2.* Percentage, with 95% CIs, of RCTs using various approaches to discuss clinical significance in clinical psychology journals ( $n = 104$  RCTs) and psychiatry journals ( $n = 89$  RCTs), in Study 1.

### **Discussion**

Almost all RCTs used significance tests to report results, but other essential information, such as a power estimate, effect size estimates and CIs, was often absent. Only 13% of RCTs included the information necessary to answer all three key questions outlined in our introduction. This is a very low proportion for studies that provide the basis of empirically-supported treatments, and that appeared in prominent and influential journals.

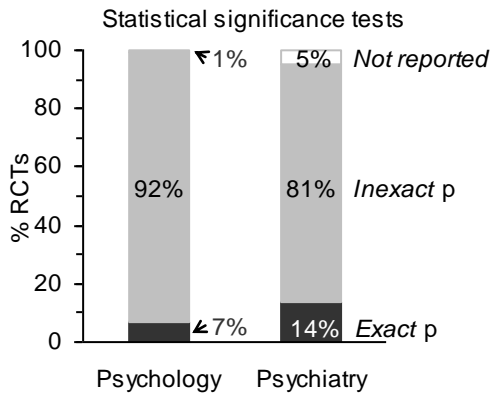
RCTs tended to have research designs of 2 (independent groups) by 4 (repeated measures), for around 10 dependent variables. Such complex designs are difficult, but perhaps not impossible to represent using graphical CIs (Bloin & Riopelle, 2005; Masson & Loftus, 2003). Less than half (48%) of clinical psychology RCTs and 60% of psychiatry RCTs presented any results at all graphically.

Compared with past research (see Table 1), the present study of RCTs found somewhat higher rates of reporting standardized or units-free effect size, and of interpreting effect size, discussing clinical significance, using figures, using significance tests, and mentioning statistical power, at least implicitly. These differences may be due partly to changes over time, partly to the

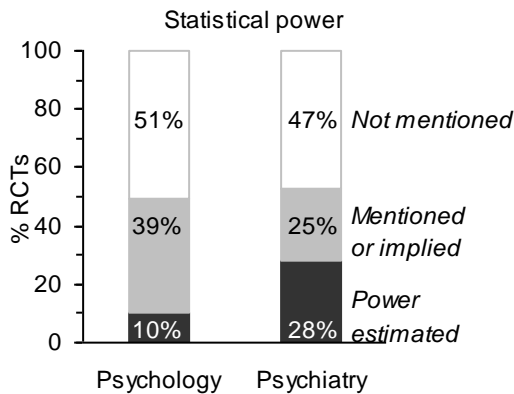
relatively rigorous nature of RCTs compared with other psychological research, and partly to the immediate clinical and practical utility of RCTs.

Even though psychiatry RCTs reported CIs much more often than psychology RCTs, the reporting was usually numerical rather than graphical, and without interpretation. Effect sizes were also reported frequently but rarely interpreted. We conclude therefore that some reporting practices have changed in psychiatry but only in a mechanical sense—resulting in greater reporting of these statistics, but little meaningful use of them.

A.



B.



*Figure 3. Study 1. A. Percentage of RCTs reporting exact and inexact  $p$  values in clinical psychology journals ( $n = 104$  RCTs) and psychiatry journals ( $n = 89$  RCTs). B. Percentage of RCTs reporting a power estimate, or mentioning power, or the possible effect of sample size on the ability to detect differences using statistical significance tests, in clinical psychology journals ( $n = 104$  RCTs) and psychiatry journals ( $n = 89$  RCTs).*

### Study 2: Survey of Authors of RCTs

Study 1 revealed poor statistical reporting practices in RCTs of psychological therapies. In Study 2, we investigated authors' opinions about which statistical practices are relevant to the purposes of RCTs.

## **Method**

### *Participants*

Email addresses were sought for all contact authors of the RCTs included in Study 1, and were found for 123 potential participants. Approximately 20% of email addresses were updated using internet and database searches. Potential participants were invited to complete a brief survey included in an email. A follow-up email was sent two months later to non-respondents. The response rate was 38% (6 incorrect addresses, 6 authors on leave, 111 emails assumed received, 42 replies).

### *Email Survey*

The first survey question was open-ended, “Think of a clinical trial that you designed. What was the most central question(s) it was designed to answer?” This was designed to access participants’ immediate understanding of an RCT’s purpose, avoiding simplified or unfamiliar examples. The response format was designed to elicit minimal, key words and appear easy to answer, to increase response rates.

The second survey question provided a series of statements, to which participants rated their agreement (1=strongly disagree, 7=strongly agree):

“When I plan a clinical trial, the most important question in my mind is:

(please type a number from 1 to 7 in each box)

Whether the treatment makes a difference (whether there is an effect)

How big an impact the treatment has (the size or magnitude of the effect)

How clinically important the effects of treatment are

Other (please specify)”

The three items (above ‘other’) reflect the three questions raised earlier in this paper: Is there a true effect? How large is the effect? To what extent is the effect clinically (or practically) important? The order of the statements within the second question was counterbalanced.

A third and final question asked participants to suggest changes they would like to see in results reporting in RCTs.

## **Results**

In response to the first (open-ended) question, most participants (86%, 25 of 29 who answered this question) described the central purpose of an RCT as establishing whether there is an effect or difference. For example: “Is CBT more effective than a credible attention placebo or a wait list in treating GAD”, and “To test whether cognitive therapy for PTSD and/or self-help are effective”. Less than a third (31%, 9 of 29) included aims that could relate to the magnitude of effects, and the vast majority (8 of 9) of these were ambiguous, for example, “To determine the differential effectiveness of therapists training and ... two types of group therapy for depression”. Fewer still (10%, 3 of 29) provided aims related to clinical significance, such as, “[To find out] whether a psychological intervention would have a clinically significant positive impact...”

However, responses to question two give a very different impression of participants’ opinions about the purpose of RCTs. As shown in Figure 4, participants rated all three question types as very important in the planning of RCTs. Most of the 42 respondents gave ratings of moderately agree and strongly agree for all three questions. The high consistency in responses is indicated by the high precision (shortness) of the CIs. ‘Other’ research questions nominated

related to cost-effectiveness, practical importance, predictors, mediators and moderators of the treatment effect.

Finally, 64% (27 of 42) suggested changes they would like to see in how results are reported in RCTs. The most frequent suggestions were: more consistent reporting practices to enable comparisons across studies (8 of 27); effect sizes reported as standard practice (9 of 27); and clinical significance discussed as standard practice (9 of 27).

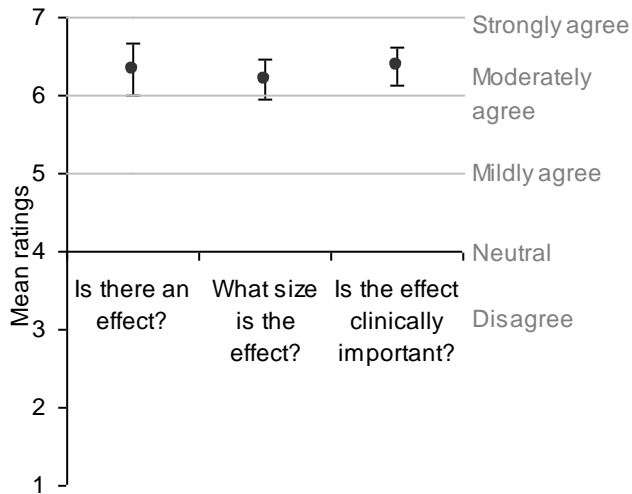


Figure 4. Mean ratings and 95% CIs for the statement, “When I plan a clinical trial, the most important question in my mind is...” ( $N = 37$ ).

### Discussion

When asked at the start to describe the main purposes of an RCT, most participants reported questions of the format, “Is there an effect of therapy?” This finding is consistent with the dominance of statistical significance tests in psychology: Research aims were—possibly unconsciously—reduced to the type of dichotomous question answerable by statistical significance tests. By contrast, when asked specifically about each of the three key questions (i.e., Is there a true effect? How large is the effect? To what extent is the effect clinically (or practically) important? ), authors rated all three as very important.

The response rate of 38% was perhaps moderate. It is likely, however, that the self-selected respondents were if anything more statistically aware and confident than non-respondents, and so our conclusions may if anything underestimate the extent to which RCT authors are focused primarily on dichotomous decision making.

When asked for further suggestions, several authors expressed a desire to compare outcomes across studies, and frustration with the difficulty in doing this with current statistical approaches. Use of standardized effect sizes would allow comparisons across measures and studies, without any need for all researchers to use the same measure for a particular variable, and as previously mentioned, CIs may facilitate meta-analytic thinking. Further, CIs give

information about true (population) effects, rather than merely sample statistics, allowing authors to present answers to all three key questions.

### **General Discussion**

There was a notable discrepancy between authors' ratings of importance in Study 2, and their reporting practices in Study 1: 86% of participants in Study 2 agreed that all three key questions were very important in RCTs, yet only 13% of RCTs in Study 1 gave information needed for all questions. Previous studies have found similar outcomes (e.g., Fidler et al., 2005; Finch et al., 2004). It seems that “authors fail to realize how often their final data presentations and interpretations reduce to NHST [significance tests]” (Finch et al., p. 321).

Study 1 included all RCTs that assessed psychological therapies for psychological disorders, and that were published during 1999-2003 in leading general psychology and psychiatry journals. Our study therefore included a considerable proportion of the best published literature that underpins evidence-based practice in clinical psychology. It is thus an important and disappointing finding that so few of the published reports provided sufficient information to answer all three key questions.

### ***Recommended Approaches to Reporting Results of RCTs***

We recommend the following approaches to reporting the results of RCTs, based on recent research, CONSORT, and the APA Task Force on Statistical Inference (CONSORT, n.d.; Faulkner, Fidler, & Cumming, 2006a; Wilkinson & Task Force on Statistical Inference, 1999).

First, instead of using statistical significance tests to guide interpretation of results, authors and readers should focus on estimating the magnitude of effects and the associated uncertainty. The most important finding is not that the therapy worked (almost all of them do), but an estimate of how well it worked.

Second, authors should use CIs to indicate the precision of the effect size estimate. Effect sizes (or means, or percentages) on their own merely tell us the size of the effect for this sample; CIs give us information about the effect size in the population. That is, CIs are inferential statistics<sup>2</sup>.

Third, effect sizes and CIs need to be interpreted, when reported. CIs offer rich information, as we explain above, but are too often reduced to statistical significance in discussions. Cumming and Finch (2005) suggested guidelines on how to substantively interpret CIs.

Fourth, where possible use figures to report CIs for main outcomes. CONSORT (n.d.) suggested reporting CIs in tables, but we disagree. CIs contain several elements of important information, including the precision of the estimate, indicated by interval width; the relation between the interval and important reference points—for example zero effect, or smallest clinically meaningful effect; and the proportion overlap of two or more intervals. Judgments about relative distances or proportions are easy to make visually from a figure, but difficult to make from numerical values reported in text or tables. CIs should be expressed on whatever measurement scale—standardised or not—best allows interpretation of effect size and clinical significance. We admit it is difficult to design figures with CIs for complex designs with many dependent variables, but hope that improved graphical designs can widen the scope for such figures (Faulkner, Fidler, & Cumming, 2006a).

Fifth, report and discuss clinical significance and the practical interpretation of the data. Does the therapy lead to recovery, or clinically important changes, and if so, in what percentage

of clients? For guidance on calculating and interpreting clinical significance, see the special section in the 1999 edition of *Journal of Consulting and Clinical Psychology* (e.g., Kazdin, 1999).

Sixth, present or refer to a meta-analysis to consolidate research. In most areas of clinical psychology research, a precise estimate (e.g.,  $\pm 2$  SD units) of the effects of an intervention can only be found through pooling the results of several studies. Don't just tell readers what your study found; tell them where your study has left the current state of research in your area.

Specific guidance for reporting results of RCTs is provided—with a worked example—in Fidler, Faulkner and Cumming (2007). Finally and most importantly we cite the recommendation of Wilkinson & Task Force on Statistical Inference (1999) for thoughtful interpretation of data. Researchers must ask themselves, what is the real meaning of these data, and how can this be best communicated to readers?

### References

- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Bezeau, S. & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, 23, 399-406.
- Bloin, D. C., & Riopelle, A. J. (2005). On confidence intervals for within-subjects designs. *Psychological Methods*, 10(4), 397-412.
- CONSORT. (n.d.). CONSORT statement. Retrieved May, 2005, from <http://www.consort-statement.org/>
- 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, 532-74.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170-180.
- Crosby, R.D., Wonderlich, S.A., Mitchell, J.E., de Zwaan, M., Engel, S.G., Connolly, K. et al (2006). An empirical analysis of eating disorders and anxiety disorders publications (1980-2000) - Part II: Statistical hypothesis testing. *International Journal of Eating Disorders*, 39, 49-54.
- Faulkner, C., Fidler, F., & Cumming, G. (2007a). *Psychologists find confidence intervals easier and more informative than significance tests, for results of randomized controlled trials*: Manuscript in preparation.
- Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics*, 33(5), 615-631.
- Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., et al. (2005). Toward improved statistical reporting in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting & Clinical Psychology*, 73(1), 136-143.
- Fidler, F., Faulkner, C. & Cumming, G. (2008). Analyzing and presenting outcomes. In A.M. Nezu & C.M. Nezu (Eds). *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions*. (pp. 315-334). New York: Oxford University Press.

- Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments & Computers*, 36(2), 312-324.
- Finch, S., Cumming, G. & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational & Psychological Measurement*, 61, 181-210.
- Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684-704.
- Harlow, L.L. (1997). Significance Testing in Introduction and Overview. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds). *What If There Were No Significance Tests?* (pp.1-17). Mahwah, NJ, USA: Lawrence Erlbaum.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
- Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting & Clinical Psychology*, 67(3), 332-339.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational & Psychological Measurement*, 56, 746-759.
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61, 213-218.
- Luborsky, L., Rosenthal, R., Diger, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., et al. (2002). The dodo bird verdict is alive and well—mostly. *Clinical psychology: Science and Practice*, 9(1), 2-12.
- Masson, M., & Loftus, G. R. (2003). Using confidence intervals for graphically based interpretation. *Canadian Journal of Experimental Psychology*, 57, 203-220.
- Newcombe, R. G., & Altman, D. G. (2000). Proportions and their differences. In D. G. Altman, D. Machin, T. N. Bryant & M. J. Gardner (Eds.), *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.). Bristol, UK: BMJ Books.
- Robbins, C. J. (1988). Attributions and depression: Why is the literature so inconsistent? *Journal of Personality and Social Psychology*, 54, 880-889.
- Rossi, J. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175-198). Mahwah, NJ: Lawrence Erlbaum.
- Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.
- Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, 92(3), 581-604.
- Smith, L. D., Best, L. A., Stubbs, A., Archibald, A. B., & Robertson-Nay, R. (2002). Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist*, 57(10), 749-761.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-760.
- 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.

- Vacha-Haase, T. & Ness, C.N. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and Practice. *Professional Psychology: Research and Practice*, 30, 104-105.
- Vacha-Haase, T. & Nilsson, J.E. (1998). Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling and Development*, 31, 46-57.
- Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes". *Psychological Bulletin*, 122(3), 203-215.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

### Footnotes

1. A difference of  $d = 0.2$  might be considered of minimal importance, and, arguably, not worth much research investment, especially given that there is still relatively little known about general therapy factors that cause larger therapy effects. However, there is little point in conducting a study where statistical non-significance is almost certain (given very low power), and then focusing only on statistical significance. At this point advocates of statistical reform raise their arms and say, "Of course it was non-significant, what did you expect?" What may be of more interest is that the difference was small, and/or clinically trivial, which is presumably what the researchers are trying to get at when they say the difference was non-significant.

2. CIs are sometimes referred to as descriptive (sample) statistics. However, CIs are inferential statistics based on similar assumptions and calculations as  $p$ -values (see footnote 1 regarding some limits to this approach). CIs estimate the magnitude of effects in the population, based on the sample data.

### Author Notes

Cathy Faulkner, Fiona Fidler and Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Australia.

Preparation of this article was supported by a Postgraduate Writing-Up Award from the Institute of Advanced Study, La Trobe University.

Correspondence concerning this article should be addressed to Geoff Cumming, School of Psychological Science, La Trobe University, Victoria 3086, Australia. Email: [g.cumming@latrobe.edu.au](mailto:g.cumming@latrobe.edu.au).