Running Head: ACCURACY OF PUBLISHED EFFECT SIZE ESTIMATES

Accuracy of Effect Size Estimates from Published Psychological Research

Andrew Brand

King's College London Institute of Psychiatry

Michael T. Bradley, Lisa A. Best and George Stoica

University of New Brunswick

Please send correspondence to:
Dr. M. T. Bradley
Department of Psychology
P.O. Box 5050
University of New Brunswick
Saint John, New Brunswick
E2L 4L5, Canada

Brand, A., Bradley, M. T., Best L. A., & Stoica, G. (2008). Accuracy of Effect Size Estimates from Published Psychological Research. *Perceptual and Motor Skills, 106,* 645-649.

An electronic version of this paper can be found at

http://ammonsscientific.com/link.php?N=44636

Summary.-A Monte-Carlo simulation was used to model the biasing of effect sizes in published studies. The findings from the simulation indicate that when a predominant bias to publish studies that have statistically significant results is coupled with inadequate statistical power there will be an overestimation of effect sizes. The consequences that such an effect size overestimation will then have on meta-analyses and power analysis are highlighted and discussed along with measures that can be taken to reduce the problem.

How Accurate are Effect Size Estimates from Published Psychological Research?

As a result of the recommendations of Wilkinson and the APA Task Force on Statistical Inference (1999) reporting effect sizes and the meta-analytic thinking it facilitates is becoming more widespread practice. The value of this practice depends on the effect size estimates approximating the true underlying effect sizes. Schmidt (1992) demonstrated that effect sizes estimates from studies that have obtained statistically significant results ($p < 0.05$) will tend to overestimate the true effect size. Moreover, Kraemer, Gardner, Brooks and Yesavage (1998) have shown that this is especially the case if the studies are statistically underpowered. Given that published psychological studies are typically statistically underpowered (Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Rossi, 1990) and there is a predominant longstanding bias to publish only studies that obtain statistically significant results (Sterling, 1959), published effect size estimates tend to overestimate the true effect size.

Although it seems that researchers have not demonstrated strong awareness of this issue, it has been investigated by quantitative psychologists (i.e., Lane & Dunlap, 1978) and its consequences have been labelled and well described (Rosenthal, 1979). Thus, it is probable that many researchers in psychology might know that published effect size estimates tend to overestimate true effect sizes but they may not fully appreciate the magnitude of the overestimations or the likely frequency of such overestimations. In this paper we conducted a Monte-Carlo simulation to show the extent to which published effect sizes typically overestimate the true effect size when the true effect size is small ($d = .20$), medium ($d = .50$) and large ($d = .80$) and what percentage of the time these various overestimations will occur when a study is conducted.

A control distribution with a standard normal distribution of 1,000,000 values and a mean of 10 and a standard deviation of 2 was created. This was accomplished by randomly generating 1,000,000 numbers between 0 and 1 then transforming these numbers into z scores using the inverse normal distribution function (see Brophy, 1985). Three experimental distributions were created so that the difference between the control distribution and an experimental distribution corresponded to one of Cohen's (1977) definitions of small (d = .20), medium (d = .50) and large (d = .80) effect sizes. This was achieved by adding a constant to every value in the control distribution to create an experimental distribution for each of the three effect size definitions. The means for the small, medium and large effect sizes were 10.40, 11.00, and 11.60, respectively and the standard deviations of the experimental distributions were the same as the control distribution (i.e., 2).

For each of the three effect sizes, 100,000 experiments were simulated by randomly sampling 38 values from the control distribution and 38 values from one of the experimental distributions. The overall sample size of 76 for a simulated experiment was derived from a survey of the statistical power of published psychological research conducted by Rossi (1990). According to Maxwell (2004), this value has been relatively constant since Cohen’s (1962) article on power and, thus, the sample size of 76 should be reasonably representative of the sample size used in a typical published psychology study with a two conditions between-subjects design. For each simulated experiment, a two-tailed between-subjects t-test was computed. Then the published effect sizes were derived. This was done by calculating and collating the observed effect sizes *only* for the

simulated experiments where the p value from the unrelated t-test was statistically significant (i.e., $p < .05$).

For each set of 100,000 simulated experiments, the mean simulated published effect sizes and the mean percentage of the differences between the simulated published effect size and the true effect size were calculated. Additionally, the probabilities of obtaining the simulated published effect sizes were computed by calculating the percentage of the simulated experiments that obtained statistical significant results ($p < .05$). These results are summarized in Table 1.

Insert Table 1 about here

The Monte-Carlo simulation therefore shows that a predominant bias to publish studies that obtain statistically significant results coupled with inadequate statistical power results in effect sizes being substantially overestimated. For instance, when a published effect size is approximately medium ($d = .58$) the true effect size could be small ($d = .20$)! The Monte-Carlo simulation also showed that the probability of a study obtaining the published effect size overestimations simply corresponds to the statistical power to obtain a statistical significant result for small, medium and large effects.

We remind the reader that we chose a single sample size that was representative of typical studies in the psychological literature. The use of this sample size emphasizes likely distortions we believe are found in the psychological research. The potential limitation in our presentation, of course, that some readers may not realize that larger sample sizes would lead to diminished distortions. We realize that we could have used a

larger range of sample sizes and effect sizes but the point is made clearly and economically with a single (and common) sample size and three common effect sizes.

Meta-analyses based on published effect size overestimates will also consequently overestimate the true effect size. One remedy would be to increase the accessibility of nonpublished research by creating a research register for all studies in an area (see Cooper, DeNeve, & Charlton, 1999). Another solution could be to employ statistical techniques. For example, subtracting the proportion of effect size variance related to N may be used to reduce the distorting effect that published effect size overestimates have upon the effect size estimates from meta-analyses (see Bradley & Stoica, 2004).

Power-analyses based on published effect size overestimates will underestimate the sample size required to obtain a statistically significant result. As a consequence, replications of an experiment are highly likely to yield statistically nonsignificant results. For instance, after conducting a power analysis based on a published medium effect size (d = .58), when the true effect size is .20, researchers may believe they have an 80% chance of detecting a statistically significant effect with a sample size of 96 participants, whereas the likelihood of actually detecting a statistically significant effect with this sample size is actually only 16%!

To prevent and minimize the published effect size overestimates, the prevailing bias to publish only research that obtains statistically significant results could be eliminated. This could be done by discontinuing statistical significance testing (Hunter, 1997 but see Abelson, 1997; Shrout, 1997) or alternatively by the adoption of blind-to-outcome peer review (Sterling, Rosenbaum, & Weinkam, 1995). Given that both of these methods would involve widespread and radical reform, increasing the statistical power of

published studies may also prove a feasible way of reducing published effect size overestimation. Increasing the statistical power of a study can be achieved by either increasing the sample size or by increasing the size of an effect. One method of increasing the sample size is to conduct web-based research (Birnbaum, 2004) but one must remember that there are biases associated with populations who have access to the web and would respond. Alternatively, the size of an effect can be increased by either reducing within-group variance (Kraemer, 1991) or by increasing the sensitivity of the dependent measure.

In summary, the Monte-Carlo simulation shows that published effect size estimates will substantially overestimate the true effect size if there is a bias against publishing nonsignificant results. Published effect size overestimation could be minimized in future psychological studies by increasing statistical power but this approach has not proven popular, as yet (Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Maxwell, 2004). Although this bias has long been recognized and a wide variety of solutions have been proposed (Cohen, 1990; Loftus, 1996; Rosenthal, 1979), the impact of recognition and solutions has not been strong. Future research might concentrate on not only attempting to find solutions to these problems but also on inducing psychologists to incorporate those solutions in their research endeavours.

References

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

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology, 55,* 803-832.

Bradley, M.T., & Stoica, G. (2004). Diagnosing estimate distortion due to significance testing in literature on detection of deception. *Perceptual and Motor Skills, 98,* 827-839.

Brophy, A. L. (1985). Approximation of the inverse normal distribution function. *Behavior Research Methods, Instruments & Computers, 17,* 415-417.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 69,* 145-153.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,* 447-452.

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

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

Kraemer, H. C. (1991). To increase power without increasing sample size. *Psychopharmacology Bulletin, 27,* 217-224.

Kraemer, H. C., Gradner, C., Brooks, J. O., & Yesavage, J.A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. Psychological Methods, 3, 23-31.

Lane, D. M, & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107-112.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in Psychological Science, 5,* 161-171.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and Remedies. *Psychological Methods, 9,* 147-163.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646-656.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in Psychology. *American Psychologist,* 47, 1173-1181.

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

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power have an Effect on the Power of Studies? *Psychological Bulletin, 105,* 309–316.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association, 54,* 30-34.

Sterling, T. D., Rosenbaum W. L., & Weinkam J. J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49,* 108-112.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Table 1. Mean Percentage of the Difference Between the Simulated Published Effect Size and the True Effect Size and the Probability of Obtaining the Simulated Published Effect Size

	True Effect Size		
	Small (d = 0.20)	Medium (d = 0.50)	Large (d = 0.80)
Mean of % of the Difference Between the Simulated Published Effect Size and the True Effect Size	+190% (0.58)	+33% (0.66)	+5% (0.84)
Probability of Obtaining the Simulated Published Effect Size	14%	58%	93%

Note. The mean for the simulated published effect sizes are in parentheses.