**Opinion** is intended to facilitate communication between reader and author and reader and reader. Comments, viewpoints or suggestions arising from published papers are welcome. Discussion and debate about important issues in ecology, e.g. theory or terminology, may also be included. Contributions should be as precise as possible and references should be kept to a minimum. A summary is not required.

# *Meta-analysis can "fail": reply to Kotiaho and Tomkins*

***Michael D. Jennions, School of Botany and Zoology, Australian National Univ., Canberra, A.C.T. 0200, Australia. (michael.jennions@ anu.edu.au). – Anders P. Møller, Laboratoire Parasitologie Evolutive, CNRS UMR 7103, Univ. Pierre et Marie Curie, Bât. A, 7ème étage, 7 quai St. Bernard, Case 237, FR-75252 Paris Cedex 05, France. – John Hunt, School of Biological, Earth & Environmental Sciences, The Univ. of New South Wales, Sydney 2052, N.S.W., Australia.***

With meta-analysis, researchers transform statistical tests of hypotheses into a common metric the 'effect size'. An effect size is 'the degree to which the phenomena is present in a population' or 'the degree to which the null hypothesis is false' (Cohen 1988, pp. 9–10). One aim of meta-analysis is to calculate the average effect size after weighting each estimate by its sampling variance. Kotiaho and Tomkins (2002) (hereafter K&T) recently suggested this procedure always yields the conclusion that the mean effect size is significantly different from zero because of strong publication bias. Their argument is based on Csada et al. (1996) who noted that in 1201 papers only 8.6% of tests of the main hypothesis concluded the effect was non-significant.

K&T (2002) illustrate their claim with an example. They assume a true mean effect of zero and that nine of every ten studies is significant due to publication bias. They then conclude that the mean effect size must be significantly greater than zero. This example is slightly misleading because it exaggerates the problem posed by publication bias. First, with a true mean of zero, significant results are equally likely to be greater or less than zero. So, in their example, even with publication bias the mean effect size calculated from published studies would be zero. One 'paradox' of publication bias, is that it is most likely to inflate the estimate of the mean effect when the true effect is small but non-zero (Palmer 2000). The real issue is thus the extent to which publication bias causes us to overestimate mean effect sizes. More generally, type I error (a significant result when the null hypothesis is correct) is only an index of the extent to which overestimation occurs. For example, would the mean effect still differ from the null hypothesis if publication bias were taken into account? Usually the null hypothesis is that the mean effect is zero, but it need not be. Second, publication bias is sensitive to both P-values and sample size (Song et al. 2000, Palmer 2000, Møller and Jennions 2001). Significant results based on small samples are published, while non-significant ones are not. With reasonable sample sizes, however, even non-significant results are eventually published. Meta-analysis gives greater weighting to studies with smaller sampling variance (i.e. larger sample sizes). In their illustrative example, K&T (2002) assumed all sample sizes were identical. This again exaggerates the effect of publication bias. Unpublished studies should have smaller sample sizes and therefore a fairly weak effect on the weighted estimate of the mean effect size.

We fully agree with K&T (2002) that publication bias is a source of concern. We disagree with their statement that "meta-analysis can not fail to provide an effect size significantly different from zero". The available data supports our perspective.

How many published meta-analyses fail to reject the null hypothesis that the mean effect is zero? We counted up the number of tests of mean effect sizes that were or were not significant in 47 published meta-analyses in biology. We simply looked at the main summary tables or figures. This was a cursory survey and we did not concern ourselves with the lack of independence between tests (e.g. we counted tests of groups A, B and C and the test of 'all' ( = A + B + C) as four tests. We also treated estimates calculated at the sample, study and species level as independent tests). Of 831 estimates of mean effect sizes, 512 were significant at the 0.05 level (62%) (Fig. 1). So, for what it is worth, meta-analyses can, and do, 'fail'.

Even if publication bias occurs, how strong an influence does it have on estimates of mean effects? Concluding that publication bias makes meta-analyses worthless is like concluding that measurement error

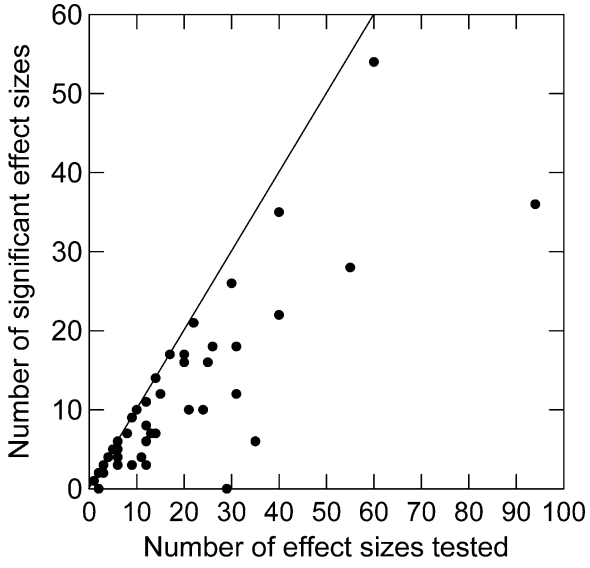


Fig. 1. The number of tests of mean effect size that yielded an estimate of the mean effect significantly different from zero ($P < 0.05$) plotted against the total number of tests performed ($N = 47$ published meta-analyses). All points should fall on the 1:1 line if meta-analyses never fail to reject the null hypothesis that the mean is zero.

makes fieldwork pointless. It depends on the magnitude of the problem. Again, we can tackle this question empirically. Jennions and Møller (2002b) examined 40 subsets of data from 40 published meta-analyses in biology. We estimated how many studies were missing using the 'trim and fill' method of Duvall and Tweedie (2000a, b). This method is based on detection of asymmetry in plots of sample size against effect size ("funnel plots"). To be conservative, we can assume studies are missing solely due to publication bias. In fact, there may be few or no unpublished studies as asymmetry in a funnel plot can occur for several other reasons (Thornhill et al. 1999). We corrected for potential publication bias by adding these 'missing' studies to the actual data sets and then recalculated the mean effect size. In 21% (8/38) of cases the weighted mean effects were no longer significantly greater than zero. Clearly this is cause for concern, however, it shows that the problem of publication bias is neither insoluble nor excessive. Rephrasing the finding, 79% of effect sizes initially estimated to be significantly greater than zero remained so even after correcting for publication bias. So these biological relationships, while overestimated, do appear to be genuine.

Even if Csada et al.'s (1996) statement that only 8.6% of main results are non-significant is correct, it is worth noting that many meta-analyses deal with data that is not the key focus of the publication. Meta-analysts know this only too well as they often struggle to track down data buried in papers asking completely different central questions. K&T (2002) argue that given only 8.6% of findings are non-significant and estimates of mean effect size rarely exceed $r = 0.3$ then, in the absence of publication bias, the true mean effect must be close to $r = 1$ for most hypotheses. They are implicitly suggesting that publication bias is extremely strong (i.e. there are far too few non-significant studies published if true effect size are less than $r = 0.3$). This argument is incomplete though. The likelihood of obtaining a significant result (i.e. the power of a test) depends on the true effect size and the sample size. They therefore need to show that the average sample sizes in biology for the tests reported by Csada et al. (1996) is such that considerably more than 8.6% of studies should report non-significant results assuming a true effect size of $r = x$. To illustrate, if Csada et al's 1201 tests all examined the significance of Pearson's correlations and the true $r = 0.50$, then only 10% (120) will fail to report a mean correlation significantly greater than zero (at alpha = 0.05, two-tailed) if the sample size per study is a modest $n = 37$. This does not differ from the observation of 103 non-significant studies ($\chi^2_1 = 1.43$, $P = 0.23$). If the average sample size is greater than 37, the true effect can be less than $r = 0.5$. For example, if the mean sample size is 113 then a true effect of $r = 0.30$ again yields only 120 studies with non-significant results. We are not disputing K&T's (2000) argument that there is probably a publication bias, only that they have overstated the problem by claiming the true effect must be close to $r = 1$ to account for the findings of Csada et al. (1996).

Finally, we would like to make two more general points. First, publication bias is a problem for any form of review, including traditional narrative reviews. Narrative summaries in individual papers or in reviews are invariably more biased than meta-analyses because the authors cannot possibly calculate effect sizes while taking sample size and heterogeneity among studies into account. Thus, while meta-analyses may be biased, we claim that narrative summaries are bound to be even more biased. Unfortunately, publication bias has become negatively linked with meta-analysis as a procedure to synthesizing the literature. We remind readers that most of the recent work looking at publication bias in biology comes from researchers who support the use of meta-analysis (Poulin 2000, Palmer 2000, Møller and Jennions 2001, Jennions and Møller 2002a, b). Earlier narrative reviews simply overlooked or ignored the problem of publication bias. Second, meta-analysis involves far more than just testing whether the mean effect differs from zero. It is also about detecting correlates of effect size and identifying factors leading to among-group differences in effect sizes since these will help to generate novel hypotheses and thereby advance science. These heterogeneity and correlational tests may be far less vulnerable to publication bias. In summary, we believe that meta-analysis, like any other scientific tool, is subject to errors and problems of application so

that results should be interpreted with caution. But even if publication bias does exist, the use of meta-analysis is still superior to traditional narrative reviewing techniques.

## References


Cohen, J. 1988. Statistical power analysis for the behavioral sciences, 2d ed. – L. Erlbaum, Hillsdale. New Jersey.

Csada, R. D., James, P. C. and Espie, R. H. M. 1996. The "file drawer problem" of non-significant results: does it apply to biological research? – Oikos 76: 591–593.

Duvall, S. and Tweedie, R. 2000a. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. – Biometrics 56: 455–463.

Duvall, S. and Tweedie, R. 2000b. A non-parametric 'trim and fill' method of assessing publication bias in meta-analysis. – J. Am. Stat. Ass. 95: 89–98.

Jennions, M. D. and Møller, A. P. 2002a. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. – Proc. R. Soc. Lond. B 269: 43–48.

Jennions, M. D. and Møller, A. P. 2002b. Publication bias in ecology and evolution: an empirical assessment using the "trim and fill" method. – Biol. Rev. 77: 211–222.

Kotiaho, J. S. and Tomkins, J. L. 2002. Meta-analysis can it ever fail? – Oikos 96: 551–553.

Møller, A. P. and Jennions, M. D. 2001. Testing and adjusting for publication bias. – Trends Ecol. Evol. 16: 580–586.

Palmer, A. R. 2000. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. – Annu. Rev. Ecol. Syst. 31: 441–480.

Poulin, R. 2000. Manipulation of host behaviour by parasites: a weakening paradigm? – Proc. R. Soc. Lond. B 267: 787–792.

Song, F., Eastwood, A. J., Gilbody, S. et al. 2000. Publication and related biases. – Health Technol. Assessment 4 (10): 1–115.

Thornhill, R., Møller, A. P. and Gangestad, S. 1999. The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. – Am. Nat. 154: 234–241.