

Publication bias in meta-analysis: seeing the wood for the trees

Joseph L. Tomkins, Div. of Environmental and Evolutionary Biology, School of Biology, Bute medical Building, Univ. of St Andrews, St Andrews, Fife, KY16 9TS UK (jtomkins@st-andrews.ac.uk). – Janne S. Kotiaho, Dept of Biological and Environmental Science, Univ. of Jyväskylä, P.O. Box 35, FIN-40351 Jyväskylä, Finland.

We thoroughly respect the recent contributions of Jennions and Møller to the appropriate use of meta-analysis in Ecology and Evolution (Møller and Jennions 2001; Jennions and Møller 2002a, b). Therefore, we are surprised to have to point out the ‘wood’ among the ‘trees’ of their own data.

Our point (Kotiaho and Tomkins 2002) was simply to raise the question: given that about 8.6% of studies report non-significant effects for the main hypothesis (Csada et al. 1996), can meta-analysis fail to find a significant overall effect size? Our conclusion was that where publication bias existed, it could not. Here we reply to the comments in Jennions et al.’s (this volume) ‘reply’ and support our original claim with evidence from a thorough empirical investigation of this subject by two of these authors (Jennions and Møller 2002a).

We illustrated our point with a hypothetical example in which a new hypothesis was proposed but which was untrue (effect size zero). This hypothetical hypothesis nevertheless attracted ten publications. We calculated that, with a publication bias causing nine out of 10 publications to be significant, that, all things being equal (i.e. methodology and sample size), there would be a significant overall effect size. Jennions et al. (this volume) make two points concerning the validity of our example. The first point is that if the hypothetical true effect size really was zero, then there would be significant positive as well as significant negative effect sizes that were published among the nine significant studies. This would of course tend to render the overall effect size calculated as non-significant, rather than significant as we calculated by assuming all of the significant results fell in the same direction. Our assumption was not without foundation however, as publication bias not only manifests itself as a bias towards significant results but also and importantly, as a bias towards *intuitive* results. This has been demonstrated a number of times recently in the form of paradigm shifts and

publication biases in Ecology and Evolution (Alatalo et al. 1997, Palmer 1999, 2000, Simmons et al. 1999, Poulin 2000, Jennions and Møller 2002a). Indeed, a directional bias is used as an indication of publication bias when examining funnel plots (Light and Pillemer 1984, Begg 1994).

Jennions et al.’s (this volume) second point is that unlike in our hypothetical study, all things are not equal, and that in reality, non-significant studies that are published are likely to have larger sample sizes than the published significant ones. This relationship between significance and sample size arises because publication bias favours the publication of small studies that are significant over small studies that are non-significant (Light and Pillemer 1984, Begg 1994). Hence by giving all of our studies the same sample size we weighted them equally in our estimation of true effect size, whereas the likelihood may be that the nine significant studies would have had smaller sample sizes and therefore be weighted less. Even so, in our example if all of the studies that were just significant ($Z = 1.96$) had a sample size of 30, the study of zero effect ($Z = 0$) would require a sample size in excess of 1914 to make the overall mean weighted Fishers z_r non-significant (using equation 4.12 on page 71 in Rosenthal 1991 and equations 18-3 and 18-8 on pages 265 and 268, respectively, in Shadish and Haddock 1994). Furthermore, a study that was just significant in the opposite (counter-intuitive) direction would require a sample size in excess of 143 which is more than four times the magnitude of the nine significant studies to make the overall weighted Fishers z_r non-significant. Hence the difference in sample size between the studies that were significant and non-significant or significant in the opposite direction would have to be extreme to invalidate our example. The extreme differences in sample sizes between significant and non-significant studies required to negate the effects that we were proposing are likely to be uncommon.

Jennions et al. (this volume) quote selectively from our paper and in doing so have made our position appear to be much more extreme than it is. For example the quote 'meta-analysis cannot fail to provide an effect size significantly different from zero' was prefaced in our original text with 'when publication bias exists' and was concerned solely with 'the significance testing of the overall effect size derived from meta-analysis'. Furthermore, we did not conclude that meta-analysis was 'worthless' as implied, neither did we claim that the problems of publication bias were 'insoluble'. What we did say was that 'where bias exists *and cannot be accounted for*, the overall effect size will be an overestimate of the true effect size'. Clearly Jennions et al. (this volume) can have no problem with this concept, as this is precisely what Jennions and Møller (2002b) show and what they highlight in their 'reply'. The fact that they found that 21% of meta-analyses had overall effect sizes that were so inflated that following the trim and fill method (Duvall and Tweedie 2000a, b) they were reduced to non-significance illustrates our point. We would not expect the trim and fill method to remove significance from all studies because not all true effect sizes are zero. Further we would not expect all data sets to be severely biased, because only high profile hypotheses are likely to be paradigm driven. Given these good reasons for the trim and fill method not to influence all data sets, we think that for 21% of meta-analyses to have their effect sizes reduced to non-significance by this technique is quite staggering, and entirely supports our view that publication bias is a serious concern. Jennions et al. (this volume) reminded readers that the recent work looking at publication bias has come from researchers who support the use of meta-analysis (Palmer 2000, Poulin 2000, Møller and Jennions 2001, Jennions and Møller 2002a, b). In recent collaborations, we too have stated that 'Meta-analysis is a useful tool to examine the generality of biological phenomena' p. 595 (Simmons et al. 1999). This was also made clear in the paper that is the subject of this reply, where we advocated the use of meta-analysis 'where publication bias has been shown not to exist'. This seems to be sound advice in the light of Jennions and Møller's (2002b) finding that '38–50% of data sets have a significant publication bias' (p. 219) and that up to 21% of studies falsely rejected the null hypothesis due to publication bias. Obviously we support the use of the trim and fill methodology (Duvall and Tweedie 2000a, b) used by Jennions and Møller (2002b) to account for publication bias and we reiterate that we made no suggestion that the problem of publication bias was insoluble.

Jennions et al. (this volume) challenge our assertion that the true mean effect size must be almost one in order for only 8.6% of studies to be non-significant in the absence of publication bias. They point out that we failed to take into account sample size and show that if

the mean sample size were a modest 37 then a true r of 0.5 would yield 10% non-significant studies. Our own simulations show that $14.7 \pm 0.2\%$ non-significant studies can arise readily by chance when the sample size is 37 and the true r is 0.5. This is nevertheless an underestimate of what the true r is *likely to be* to yield 8.6% of studies non-significant by chance alone. This is because sample size distributions in meta-analyses tend to be skewed heavily towards lower values (e.g. sex ratios; West and Sheldon 2002, skew \pm SE = 1.463 ± 0.536 , $t = 2.729$, $df = 16$, $P < 0.05$; good-genes effects in sexual selection, Møller and Alatalo 1999 = 4.452 ± 0.434 , $t = 10.25$, $df = 27$, $P < 0.001$; Fluctuating asymmetry and growth, fecundity and survival, Møller 1999 = 4.572 ± 0.306 , $t = 14.56$, $df = 59$, $P = 0.001$; fluctuating asymmetry and sexual selection, Møller and Thornhill 1998 = 2.808 ± 0.264 , $t = 10.6$, $df = 81$, $P < 0.001$). Hence, for example in Møller and Thornhill's (1998) meta-analysis of the role of FA in sexual selection the median sample size for studies was a respectable 59 (mean = 80), nevertheless, 25% of studies had sample sizes of 21 or less (20% 16 or less and 10% 12 or less). To demonstrate the effect of this skew we ran a simple simulation. We generated 10 sets of 2500 random values from a t -distribution with $n = 16$ and transformed these with a standard formula (Eq. 2.16 on page 19 in Rosenthal 1991) to r statistics. To simulate true $r = 0.5$ we added 0.5 to each of the values and calculated the proportion of studies that would be just non-significant (with $n = 16$ r should be less than 0.497). With this simulation we found that where $n = 16$ and the true $r = 0.5$, $49.7 \pm 0.3\%$ of r values are non-significant. Therefore, with true $r = 0.5$, when the distribution of sample sizes is skewed such that 20% of studies have sample sizes of 16 or less we should immediately expect in excess of 9.9% of all studies to be non-significant regardless of the mean or median sample size. These skewed distributions in meta-analysis data sets also underestimate the skew that would be evident if all data sets were published, as studies of low sample size tend not to be published, while studies of large sample size do. We acknowledge that Jennions et al. (this volume) are correct that we overestimated what the true r would have to be, if only 8.6% of studies that were actually performed, rather than reported, find non-significant results. Nevertheless, the skew towards small sample sizes means that Jennions et al's estimates of true effect sizes of $r = 0.5$ and $r = 0.3$, based on mean sample sizes of 37 and 113 respectively, are both underestimates. Furthermore, because 'in evolution and ecology mean effect sizes are weak (mean $r = 0.21$ – 0.27 ,' p. 217 in Jennions and Møller 2002a), an effect size of $r = 0.5$ is likely to be extremely rare and even an effect size of $r = 0.3$ is considerably higher than what has been found for the majority of hypotheses.

Jennions et al. (this volume) have gone to considerable effort to show that meta-analyses 'can and do fail',

finding that within 47 studies, there were 831 estimates of mean effect size, of which 38% were non-significant. However, we were not specifically concerned with the numerous small meta-analyses contained within other larger meta-analytic studies. We were only concerned with the overall effect sizes, this was because Csada et al.'s (1996) study was only concerned with main hypotheses, and these are likely to be the focus of the general effect sizes in meta-analysis. Our specific concern was that the general overall effect sizes could not fail to be significant given the small number of non-significant studies in the literature. We are grateful for Jennions et al. (this volume) for providing us with the opportunity and the data, to show that this was correct. As described in Jennions et al. (this volume), Jennions and Møller (2002b) have analysed 40 published meta-analyses for evidence of publication bias. To test our hypothesis that publication bias means that meta-analysis cannot fail we can ask how many of the 40 meta-analyses investigated by Jennions and Møller (2002b) failed to reach significance? Quoting from the first line of the *results* in Jennions and Møller (2002b) (page 215): 'of the 40 meta-analyses, the initial estimate of weighted mean effect size differed significantly from zero ($p < 0.05$) in 38' (calculated using the preferred random effects model to analyse the data, see Jennions and Møller 2002b, p. 214). Hence their data shows that only 5% of meta-analyses fail to reach overall significance. As outlined above, Jennions et al. (this volume) have empirically shown that the observed low level of acceptance of null hypotheses is due to the fact that in up to 21% of all meta-analytical studies the null hypothesis is being falsely rejected as a result of publication bias: their own evidence proves our case.

Meta-analyses carry the weight of all of the studies that they summarise. This credence makes it imperative that meta-analysis can be trusted to be an impartial tool and makes the validity of meta-analytic summary a far more important issue than measurement error in fieldwork. The trim and fill methodology, exploited so thoroughly by Jennions and Møller (2002b), provides a means to objectively assess the confidence that we can have in meta-analysis.

Acknowledgements – JLT is funded by a David Phillips Fellowship from the BBSRC and JSK is funded by the Academy of Finland.

References

- Alatalo, R. V., Mappes, J. and Elgar, M. A. 1997. Heritabilities and paradigm shifts. – *Nature* 385: 402–403.
- Begg, C. B. 1994. Publication bias. – In: Cooper, H. and Hedges, L. V. (eds), *The handbook of research synthesis*. Sage, pp. 399–410.
- 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. A non parametric “trim and fill” method of assessing publication bias in meta-analysis. – *J. Am. Stat. Assoc.* 95: 89–98.
- Duvall, S. and Tweedie, R. 2000b. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias. – *Biometrics* 56: 455–463.
- Jennions, M. D. and Møller, A. P. 2002a. Publication bias in ecology and evolution: an empirical assessment using the trim and fill method. – *Biol. Rev.* 77: 211–222.
- Jennions, M. D. and Møller, A. P. 2002b. 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., Møller, A. P. and Hunt, J. 2004. Meta-analysis can ‘fail’: reply to Kotiaho and Tomkins. – *Oikos* 104: 191–193.
- Kotiaho, J. S. and Tomkins, J. L. 2002. Meta-analysis can it ever fail? – *Oikos* 96: 551–553.
- Light, R. J. and Pillemer, D. B. 1984. *Summing up: the science of reviewing research*. – Harvard Univ. Press.
- Møller, A. P. 1999. Asymmetry as a predictor of growth, fecundity and survival. – *Ecol. Lett.* 2: 149–156.
- Møller, A. P. and Alatalo, R. V. 1999. Good-genes effects in sexual selection. – *Proc. R. Soc. Lond. B* 266: 85–91.
- Møller, A. P. and Jennions, M. D. 2001. Testing and adjusting for publication bias. – *Trends Ecol. Evol.* 16: 580–586.
- Møller, A. P. and Thornhill, R. 1998. Bilateral symmetry and sexual selection: a meta-analysis. – *Am. Nat.* 151: 174–192.
- Palmer, A. R. 1999. Detecting publication bias in meta-analyses: a case study of fluctuating asymmetry and sexual selection. – *Am. Nat.* 154: 220–233.
- 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.
- Rosenthal, R. 1991. *Meta-analytic procedures for social research*, 2nd edn. – Sage.
- Shadish, W. R. and Haddock, C. K. 1994. Combining estimates of effect size. – In: Cooper, H. and Hedges, L. W. (eds), *The handbook of research synthesis*. Sage, pp. 261–282.
- Simmons, L. W., Tomkins, J. L., Kotiaho, J. S. and Hunt, J. 1999. Fluctuating paradigm. – *Proc. R. Soc. Lond. B* 266: 593–595.
- West, S. A. and Sheldon, B. C. 2002. Constraints on the evolution of sex ratio adjustment. – *Science* 295: 1685–1688.