Biol. Rev. (2002), **77**, pp. 211–222 © *Cambridge Philosophical Society*
DOI: 10.1017/S1464793101005875 *Printed in the United Kingdom*

Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method

MICHAEL D. JENNIONS[1,2]* and ANDERS P. MØLLER[3]

[1] *School of Botany and Zoology, Australian National University, Canberra, A.C.T.* 0200, *Australia*
[2] *Smithsonian Tropical Research Institute, Unit* 0948, *APO AA* 34002-0948, *USA*
[3] *Laboratoire d'Ecologie Evolutive Parasitaire, CNRS FRE* 2365, *Universite Pierre et Marie Curie*, 7, *quai St. Bernard, Case* 237, *F*-75252 *Paris Cedex* 5, *France*

(*Received* 9 *July* 2001; *revised* 29 *October* 2001; *accepted* 20 *November* 2001)

ABSTRACT

Recent reviews of specific topics, such as the relationship between male attractiveness to females and fluctuating asymmetry or attractiveness and the expression of secondary sexual characters, suggest that publication bias might be a problem in ecology and evolution. In these cases, there is a significant negative correlation between the sample size of published studies and the magnitude or strength of the research findings (formally the 'effect size'). If all studies that are conducted are equally likely to be published, irrespective of their findings, there should not be a directional relationship between effect size and sample size; only a decrease in the variance in effect size as sample size increases due to a reduction in sampling error. One interpretation of these reports of negative correlations is that studies with small sample sizes and weaker findings (smaller effect sizes) are less likely to be published. If the biological literature is systematically biased this could undermine the attempts of reviewers to summarise actual biology relationships by inflating estimates of average effect sizes. But how common is this problem? And does it really affect the general conclusions of literature reviews? Here, we examine data sets of effect sizes extracted from 40 peer-reviewed, published meta-analyses. We estimate how many studies are missing using the newly developed 'trim and fill' method. This method uses asymmetry in plots of effect size against sample size ('funnel plots') to detect 'missing' studies. For random-effect models of meta-analysis 38% (15/40) of data sets had a significant number of 'missing' studies. After correcting for potential publication bias, 21% (8/38) of weighted mean effects were no longer significantly greater than zero, and 15% (5/34) were no longer statistically robust when we used random-effects models in a weighted meta-analysis. The mean correlation between sample size and the magnitude of standardised effect size was also significantly negative ($r_s = -0.20$, $P < 0.0001$). Individual correlations were significantly negative ($P < 0.10$) in 35% (14/40) of cases. Publication bias may therefore affect the main conclusions of at least 15–21% of meta-analyses. We suggest that future literature reviews assess the robustness of their main conclusions by correcting for potential publication bias using the 'trim and fill' method.

Key words: effect size, fail-safe number, fluctuating asymmetry, funnel plots, meta-analysis, publication bias, trim and fill.

* Author for correspondence at address 1 (e-mail: Michael.Jennions@anu.edu.au Tel: +61 2 6125 3540 Fax: +61 2 6125 5573).

CONTENTS

I. INTRODUCTION

The use of meta-analysis quantitatively to review a field of study is increasingly popular in biology (Arnqvist & Wooster, 1995; Møller & Jennions, 2001). Meta-analysis summarises the literature on a topic by transforming statistical tests of hypotheses into a common metric ('effect size'). 'Effect size' is 'the *degree* to which the phenomenon is present in a population' or 'the degree to which the null hypothesis is false' (Cohen, 1988, pp. 9–10). Meta-analysis allows for quantitative answers to questions about the average strength of an hypothesised relationship, or the extent and possible sources of heterogeneity in research findings. It has clear advantages over traditional narrative reviews but, as with any review process, it assumes that the scientific literature is unbiased. Ironically, the greater precision provided by meta-analysis has also prompted biologists to question whether the scientific literature really does accurately reflect the results of the many studies biologists initiate (Csada, James & Espie, 1996; Bauchau, 1997; Alatalo, Mappes & Elgar, 1997; Simmons *et al*., 1999; Palmer, 1999; Poulin, 2000; for the medical literature see Song & Gilbody, 1998).

There are many forms of bias in the scientific literature. Some are fairly innocuous such as preferential citation of studies supporting the author's views, or by those of the same nationality (comprehensively reviewed by Song *et al*., 2000), or even a tendency to cite more often authors with surnames beginning with letters near the start of the alphabet (Trengenza, 1997). Most troubling, however, is the situation where the magnitude and/or direction of research findings influences whether or not a completed study is submitted, positively reviewed and eventually accepted for publication. No malevolent intent to suppress findings is required to generate a 'publication bias', only a systematic prejudice at any stage of the publishing process (Palmer, 2000; Song *et al*., 2000; Møller & Jennions, 2001).

The most widely cited prejudice of researchers, reviewers and editors is towards statistically significant results (Palmer, 2000; Song *et al*., 2000; Møller & Jennions, 2001). In practice, there is probably an interaction between sample size and statistical significance. For a non-significant result to be published sample sizes must be large. This is reasonable because the statistical power to detect a significant difference is low when samples are small (Cohen, 1988) so the null hypothesis of the absence of an effect of a given magnitude (usually non-zero) can not be accepted with a reasonable degree of confidence. When a result is significant, however, reviewers and editors often ignore sample size. This is not a major problem if the true scientific relationship being examined is close to zero. It simply means there will be selective reporting of non-significant findings from studies with small samples: when the average relationship is calculated it will still be close to zero. The real problem arises when the true relationship is moderate (Palmer, 1999). For studies with small samples, the only results published will tend to be those that are significant in the direction of the true effect (very few studies with an estimated effect opposite in direction to the 'true' effect will reach significance). This can lead to a systematic overestimation of the true effect size (Begg, 1994).

Although biologists are now aware of the problem, there has been no systematic attempt to determine its extent (Møller & Jennions, 2001). Is publication bias so severe that it grossly exaggerates the biological significance of certain phenomena, even

generating 'collective illusions'? Palmer (2000) has argued that this could be the situation, but did not quantify the average effect of publication bias on general conclusions. How widespread is the problem? Ideally this question is resolved by obtaining information on completed but unpublished studies to see whether their inclusion alters the conclusions of meta-analyses. Unpublished studies are, however, extremely difficult to track down. There are several ways to try to model and even to correct for effects of publication bias (e.g. weighted distribution theory, general linear models and Bayesian modelling) (Begg, 1994; Gleser & Olkin, 1996). Unfortunately though the models developed to date are not implemented in readily available, user-friendly software; they require restrictive assumptions about the exact effects of probability values and sample sizes on publishability; and they are only accessible to those with advanced statistical modelling skills (DuMochel & Harris, 1997).

In a survey of 44 ecological and evolutionary meta-analyses, we identified a significant absence of studies with small sample sizes that present findings weaker than the weighted average effect size (Jennions & Møller, 2002). In other words, we found the mean correlation between sample size and the absolute value of the effect size to be significantly negative. Specific published examples of this phenomenon are given in the second paragraph of Section II.3. Publication bias is therefore a general phenomenon. Here, we use a new and simple method developed by Duvall & Tweedie (2000*a*, *b*) called 'trim and fill' to estimate the number of unpublished or 'missing' studies. To our knowledge, the only other study to use this approach to estimate the potential impact of publication bias is an analysis by Sutton *et al.* (2000*a*) of a set of meta-analyses of clinical medical trials. According to them, the only previous general statistical assessments of the prevalence of 'missing' studies in a collection of meta-analyses was that of Egger *et al.* (1997). We then test how robust general conclusions in biology are when 'corrected' for potential publication bias.

II. METHODS

(1) Data set

We made an extensive survey of the ecological and evolutionary literature for meta-analyses published up until the end of 2000. We examined the journals *American Naturalist*, *Animal Behaviour*, *Behavioral Ecology*, *Behavioral Ecology and Sociobiology*, *Ecological Monographs*, *Ecology*, *Evolution*, *Evolutionary Biology*, *Journal of Evolutionary Biology*, *and Quarterly Review of Biology*. We also entered the phrase 'meta-analy*' into the electronic database 'WebSpiris' to find papers where this term occurred in the title or abstract. We then examined the title and place of publication of each paper listed and directly inspected any that seemed related to evolutionary or ecological biology. Furthermore, we contacted a number of colleagues who have used meta-analyses in their research to locate meta-analyses currently in press. We excluded meta-analyses of genetic heritabilities because it is unclear whether h^2 itself or an effect size based on the strength (rather than slope) of the phenotypic relationship between relatives is the more appropriate effect size. Palmer (2000) has already shown that the problem of publication bias is especially severe for h^2 because negative values are biologically irrelevant and therefore under-reported. We found 40 peer-reviewed meta-analyses where we could also obtain effect sizes and variances for the original studies (either because they were included as appendices in the published paper or where the authors kindly responded to our request and sent us the data). The ability to detect unpublished studies relies on the asymmetric distribution of effect sizes (see below). Such asymmetry is unlikely to be detected with smaller samples. We therefore set a minimum sample size of eight studies per meta-analysis (Sutton *et al.*, 2000*a* used a minimum of 10). This only removed one otherwise usable meta-analysis (Fernandez-Duque & Valeggia, 1994). The meta-analyses we used are listed in Section VIII.

Most of the 40 original meta-analyses asked several different questions (i.e. examined several response variables). In such cases, different response variables were often taken from the same or a closely overlapping set of original empirical studies. To be statistically conservative, we limited our analysis to one response variable per original meta-analysis (that with the largest sample size). In addition, the original authors often found significantly more heterogeneity in effect size among studies than could be explained by sampling error. They therefore looked for underlying structure in the data by classifying studies into groups (e.g. birds *versus* insects) and testing for significant among-group variance in effect sizes for each categorical factor using Q_b (Q_b is a measure of the variance in effect size accounted for by differences among groups) (Rosenberg, Adams & Gurevitch, 2000). For each of the original meta-analyses we therefore split the data using whichever categorical factor generated the

greatest differences in effect sizes among groups (but only if $P<0.05$ for Q_b). Finally, using these criteria we selected the group with the largest number of empirical studies (mean $\pm$ s.e.m. $= 45.1 \pm 6.97$, range $= 8$–246). If there were two or more groups with an equal number of studies we picked one at random. We used the same effect size type as the original authors. These were Pearson's r ($N = 21$), Hedges' d ($N = 10$), the natural log of the response ratio ($\ln RR$) ($N = 7$) or a customised effect size ($N = 2$).

(2) Calculating mean effect sizes

We calculated mean effect sizes weighted for sample size using the software package Metawin 2.0 (Rosenberg *et al.*, 2000). We ran both fixed-effect (FE) and random-effect (RE) models for each of the 40 data sets. FE models assume a single true effect common to all the studies. Variation in the observed effects is solely attributed to sampling error. RE models allow for a true random component as a source of variation in effect size between studies, as well as sampling error. In general, RE models are preferred (N.R.C., 1992), especially in biology where there is almost certainly real variation in actual effect sizes among different taxa or ecosystems (Gurevitch & Hedges, 1999). Some earlier meta-analyses, however, only used FE models. Our estimates of weighted mean effect sizes may differ slightly from those reported in the original papers because of rounding errors and/or minor differences in the coding of original studies. We must stress that our main intent is not to criticise individual studies, but to highlight wider trends. We then used bootstrapping with 999 replications to calculate bias-corrected 95% confidence intervals. This does not require that effect sizes are parametrically distributed. The weighted mean effect is significantly different from zero if the 95% confidence intervals do not overlap zero. Analyses based solely on parametric confidence intervals yielded the same conclusions in 139 out of 146 cases.

(3) Testing for publication bias

A funnel plot of effect size against log-transformed sample size should produce a funnel shape symmetric around the 'true' effect size (Light & Pillemer, 1984). Purely due to sampling error (the larger the sample the more accurate the estimate) the variance in estimates of the 'true' effect size is higher for studies with smaller samples. The observed effect sizes should be normally distributed around the mean effect with no trend in relation to sample size (Light & Pillemer, 1984; Begg, 1994). If studies with statistically significant results are more likely to be published, however, and the true mean is close to zero, this will produce a 'hollowed out' funnel (see Palmer, 2000). If the true effect is moderate and non-significant results tend not to be published, this will produce a skewed funnel in which the magnitude of the effect size decreases as sample size increases (Begg & Mazumdar, 1994; Palmer, 1999). This second publication bias will lead to an inaccurate estimate of the true effect size. Of course, a skewed funnel plot can be caused by factors other than publication bias since prior knowledge of effect sizes from pilot studies, reduced sample sizes for certain species, choice of effect measures, chance and many other confounding variables may also create asymmetric plots (Thornhill, Møller & Gangestad, 1999; Gurevitch & Hedges, 1999). Even so, the robustness of meta-analytic conclusions can be tested by making the conservative assumption that skew is due to publication bias.

In the biological sciences, aside from the fail-safe number (see below), the only statistical approach widely used to test for publication bias has been to test for a significant relationship between sample size and effect size using rank correlation tests (Begg & Mazumdar, 1994; for related approaches see Macaskill, Walter & Irwig, 2001; Møller & Jennions, 2001). Palmer (1999) has called this correlation r_{bias}. If studies with small samples are only published when they have significant results, and the 'true' effect is moderate, the funnel plot will be skewed. There will be a decline in the magnitude of the effect size with increasing sample size because for studies with small sample sizes there is less likelihood that an effect opposite in magnitude to the 'true' effect will reach statistical significance. This decline has now been reported in a few specific fields of study (e.g. Palmer, 1999, 2000; Gontard-Danek & Møller, 1999; Jennions, Møller & Petrie, 2001). Recently, using 232 data sets from 44 evolutionary ecology meta-analyses that included most of the current database, we found that the average relationship is a highly significant, but small, decline in effect size with sample size (Jennions & Møller, 2002). Here, we estimate these correlations using Spearman's r specifically to determine whether r_{bias} and the estimated number of studies missing based on funnel plot asymmetry (see below) are related. Although several authors, including ourselves, have previously presented the correlation between sample size and effect size (e.g. Palmer, 2000; Jennions *et al.*, 2001), strictly speaking, effect size should first be

standardised to conform to the assumptions of the test (see Begg & Mazumdar, 1994). We therefore standardised effect size here, although this makes little difference in most cases (M. D. Jennions & A. P. Møller, unpublished data).

More recently, the funnel plot has been used to derive a non-parametric method of testing and adjusting for possible publication bias in meta-analysis. The 'trim and fill' method of Duvall & Tweedie (2000*a*, *b*) estimates the number of 'missing' studies due to publication bias. The method is reliant on the symmetric distribution of effect sizes around the 'true' effect size if there is no publication bias, and the simple assumption that the most extreme results have not been published. These will usually be studies with smaller sample sizes, because variance in effect size (hence extreme values) increases as sample size decreases. Once the number of 'missing' studies is estimated, one recalculates the weighted mean effect size and its variance when they are incorporated.

The statistical procedure involves an iterative process (Duvall & Tweedie, 2000*a*, *b*). To start, one calculates the weighted mean effect size for the full data set and then 'trims off' the outlying part of the funnel plot that is asymmetrical with respect to the mean. Simple formulae are used to estimate the number of studies in the asymmetric part. These studies are then temporarily removed ('trimmed') and the remainder used to re-estimate the weighted mean effect. Then, again using the full set of studies, one 'trims off' those studies asymmetrical with respect to the new estimate of the mean. After just a few iterations the estimate of the number of studies that need to be trimmed reaches an asymptotic value. One can now 'fill in' the 'missing' studies. These are simply the mirror-image counterparts of the trimmed studies around the final weighted mean effect estimated using the symmetric portion of the data set. The missing studies are given the same variance as their corresponding 'trimmed' counterparts. Finally, the full data set that includes the trimmed, missing and remaining studies is used to calculate the new mean effect size and its confidence intervals.

Duval & Tweedie (2000*a*, *b*) present three different estimators (R_0, L_0, Q_0) for the number k of missing studies. Of these, L_0 is the best general-purpose estimator. Here, we follow Sutton *et al.* (2000*a*) in using L_0 to calculate k using formulae entered onto an Excel spreadsheet (freely available on request). We did this for both FE and RE models in case the choice of model has an impact on the assessment of publication bias. However, because RE models are more appropriate we place greater emphasis on the results for these models (N.R.C., 1992). Of course, chance asymmetry in a funnel plot will lead to positive values of L_0 (Sterne, 2000). We therefore also estimated how many meta-analyses had a value of L_0 that was significant at the 0.05 level (i.e. more studies missing than expected by chance). Critical values of L_0 were estimated by extrapolating from the simulations in Table 4 of Duvall & Tweedie (2000*b*). These are therefore crude approximations. Finally, we also calculated the fail-safe number of studies needed to nullify an effect at the 5% level following Rosenthal (1991, p. 104). This number estimates how many studies with a mean effect of zero are needed to change a significant effect to a non-significant one at the stated P level (here significance must be calculated using parametric 95% confidence intervals). A value of X greater than $5K+10$ is usually considered to indicate a robust conclusion, where K is the reported number of studies. Unless otherwise stated data are presented as the mean ± S.E.M. For non-significant tests, we present the statistical power to detect a medium effect as defined by Cohen (1988) with P (two-tailed) = 0.05.

III. RESULTS

Of the 40 meta-analyses, the initial estimate of weighted mean effect size differed significantly from zero ($P<0.05$) in 38 RE models and 35 FE models. There were only three cases in which the conclusion differed depending on the choice of model. The weighted mean effects estimated using RE models were, on average, 29% greater than those from FE models (Wilcoxon's test, $z = 3.92$, $N = 40$, $P = 0.0001$).

With RE models, 30 out of 40 meta-analyses were estimated to have one or more missing studies ($L_0>0$); with FE models this rose to 36 meta-analyses. We then noted whether the probability of each observed L_0 was less than 0.05. For RE models, 15 meta-analyses showed a significant publication bias; for FE models, 20 were significant. Thus, 38–50% of published meta-analyses show strong evidence for publication bias. The estimated number of studies missing was positively correlated for the 36 meta-analyses where both model types indicated that the studies were missing from the same side of the distribution (or where no study was missing for one or both models) ($r = 0.429$, $P = 0.009$, $N = 36$). There were, however, four cases where the side of the distribution from which the studies were missing

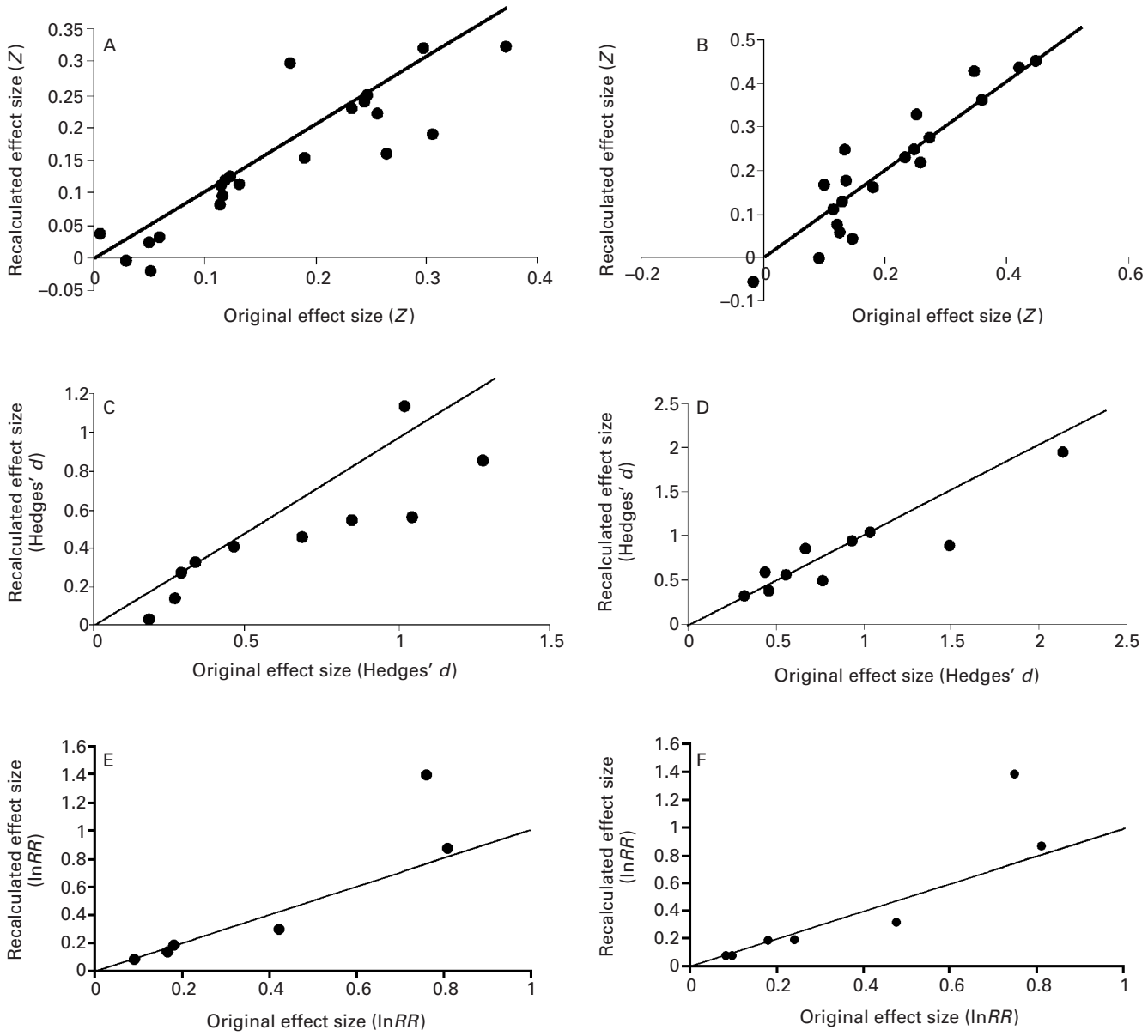

Fig. 1. Original *versus* recalculated effect sizes for: (A) fixed-effects (FE) models for effect type Z-transformed r ($Z_{new} = -0.004 + 0.900Z$; $r^2_{adj} = 77.8\,\%$, $F_{1,19} = 71.1$, $P < 0.0001$, $N = 21$; (B) random-effects (RE) models for Z ($Z_{new} = -0.0261 + 1.120Z$; $r^2_{adj} = 85.3\,\%$, $F_{1,19} = 116.7$, $P < 0.0001$, $N = 21$); (C) FE models for effect type Hedges' d ($d_{new} = 0.008 + 0.731d$; $r^2_{adj} = 72.1\,\%$, $F_{1,8} = 24.2$, $P = 0.0012$, $N = 10$); (D) RE models for Hedges' d ($d_{new} = 0.125 + 0.763d$; $r^2_{adj} = 81.1\,\%$, $F_{1,8} = 39.6$, $P = 0.0002$, $N = 10$); (E) FE models for effect type natural log of the response ratio ($\ln RR_{new} = -0.110 + 1.463 \ln RR$; $r^2_{adj} = 76.9\,\%$, $F_{1,5} = 21.0$, $P = 0.006$, $N = 7$); (F) RE models for $\ln RR$ ($\ln RR_{new} = -0.095 + 1.468 \ln RR$; $r^2_{adj} = 80.7\,\%$, $F_{1,5} = 26.1$, $P = 0.004$, $N = 7$). In all graphs, the line of equality is shown.

differed between FE and RE models. On average, the estimate of the number of missing studies was higher for RE models than for FE models (sign test, $P = 0.031$; 7.1 ± 1.2 *versus* 5.8 ± 1.1).

For FE models in 75 % (27/36) of cases where we estimated that studies were missing, their addition moved the mean effect closer to zero. There was thus a significant trend for missing studies to make conclusions about the weighted mean effect less robust (binomial test, $P < 0.005$). For RE models, however, in only 57 % (17/30) of cases where studies were missing did their addition move the mean effect closer to zero. Correction for missing studies was therefore equally likely to reduce or increase the robustness of conclusions about the weighted mean effect (binomial test, $P = 0.58$; power: 36 %). Once

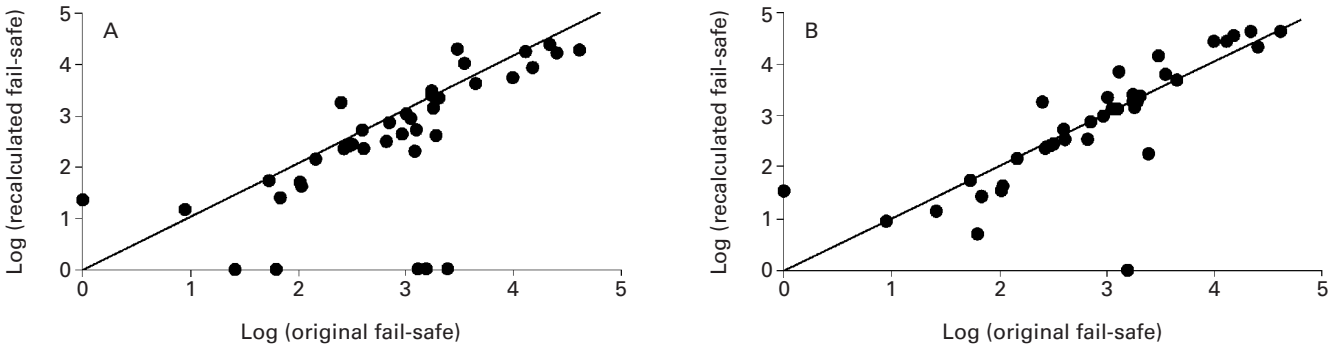

Fig. 2. Log_{10} fail-safe number (X) before and after recalculation for: (A) fixed-effects models ($X_{\text{new}} = -0.0083 + 0.8778X$; ($r^2_{\text{adj}} = 41.0\,\%$, $F_{1,38} = 28.1$, $P < 0.0001$); (B) mixed-effects models ($X_{\text{new}} = 0.0501 + 0.965X$; ($r^2_{\text{adj}} = 63.8\,\%$, $F_{1,38} = 69.7$, $P < 0.0001$) (both $N = 40$).

we had added missing cases, we recalculated effect sizes to see if this changed the conclusions of the original meta-analyses based on 95 % confidence intervals. For FE models, in three cases an initially significant result became non-significant. For RE models, in eight cases a significant result became non-significant. (These 11 cases were from 11 different meta-analyses.) No means that were originally judged non-significant became significant. In sum, 8–21 % of published meta-analyses may have been interpreted incorrectly.

The potential for publication bias was also reflected in the correlation between effect size variance and the magnitude of the observed effect size (standardised). The mean Begg–Mazumdar correlation was $r_s = -0.201$ ($t = 5.34$, $d.f. = 39$, $P < 0.0001$). Thus, the effect size was closer to zero as sample size increased. There were 14 significantly negative correlations (at $P < 0.10$), but no significantly positive correlations. [Begg & Mazumdar (1994) recommended the use of $P = 0.10$ because of the low power of the test.] We then tested if the Begg–Mazumdar correlation predicts the number of studies that are missing according to 'trim and fill' methods and whether these studies are greater or smaller than the weighted mean effect. We coded L_0 as positive if the addition of missing studies decreased the mean effect. When the Begg–Mazumdar r_s is negative, this implies that studies with effects smaller than the weighted mean are missing so L_0 should be positive. We therefore predict that L_0 and Begg–Mazumdar r_s are negatively correlated. Because the absolute value of L_0 will increase with sample size by chance, we first divided L_0 by the number of studies in the meta-analysis. As predicted there was a significant negative correlation for both FE models ($r = -0.423$, $P = 0.007$, $N = 40$) and RE models ($r = -0.397$, $P = 0.011$, $N = 40$).

Of course, a change from a significant to a non-significant weighted mean effect is a crude measure of the importance of missing studies if the real aim is to see how robust the original estimate of effect size is to publication bias. We examined this by looking at the three main effect types separately. For Pearson's r, the mean percentage change (calculated as the absolute difference in r before and after recalculation divided by the original value of $|r|$) was 57.4 ± 27.1 % for FE models and 36.5 ± 11.8 % for RE models. This corresponded to mean absolute changes in r of 0.037 ± 0.008 and 0.039 ± 0.008, respectively ($N = 21$). For Hedges' d the mean percentage change was 30.3 ± 7.9 % for FE models and 17.0 ± 5.4 % for RE models. This corresponded to absolute changes in d of 0.187 ± 0.053 and 0.153 ± 0.061, respectively ($N = 10$). For the natural log of the response ratio ($\ln RR$) the mean percentage change was 20.9 ± 10.9 % for FE models and 27.0 ± 10.2 % for RE models. This corresponded to absolute changes in $\ln RR$ of 0.121 ± 0.086 and 0.135 ± 0.085, respectively ($N = 7$). The percentage change is large even though the absolute difference in effect size estimates is small because in evolution and ecology mean effect sizes are weak (mean $r =$ 0.21–0.27; A. P. Møller & M. D. Jennions, in preparation). Original and recalculated effect sizes were similar, but 15–28 % of the variation in recalculated mean effect size is unexplained by the original estimate of effect size (Fig. 1).

When results are non-significant, researchers often use an *a priori* estimate of effect size to calculate power. In the absence of pilot studies, they may rely on a general estimate of effect sizes for relationships in their field of study (e.g. A. P. Møller & M. D. Jennions, in preparation). Most researchers, however, provide power for statistical tests assuming that the true effect is small, medium or large as defined by

Cohen (1988) (e.g. small: $r = 0.1$, $d = 0.2$; medium: $r = 0.2$, $d = 0.5$; large $r = 0.3$, $d = 0.8$). We therefore classified effect sizes before and after recalculation using the criteria of Cohen (1988). We could only do this for 31 meta-analyses (those using r or Hedges' d). For RE models, the weighted mean effect remain unchanged and was classified as large for seven, medium for 10 and small for 11 estimates. One large estimate became medium, and one medium estimate became small; while one small estimate became medium. Thus, 10 % (3/31) of effect sizes had to be reclassified after correcting for publication bias.

Finally, we looked at the robustness of weighted mean effects by examining fail-safe numbers. By the convention that $X > (5K + 10)$ is robust, 85 % (34/40) of the original weighted mean effect estimates were robust. With FE models, 20.6 % (7/34) of the estimates were no longer robust after being recalculated. For RE models, 14.7 % (5/34) of the effect sizes were no longer robust after being recalculated. For the 34 original results that were robust, in 11 cases for FE models (32 %) and in 21 cases for RE models (62 %), the weighted mean effect was the same or more robust after recalculation (Fisher's Exact test, $P = 0.028$) (Fig. 2A, B).

IV. DISCUSSION

To start, there was broad agreement between FE and RE models as to whether or not weighted mean effects were significant, although estimates from RE models were approximately 29 % larger. (The larger mean estimates for RE models may 'compensate' for the generally broader confidence intervals for RE models which, all else being equal, should decrease the likelihood of reporting a significant effect size with RE models.) If anything, the initial use of FE models by biologists may have led to weighted mean effect sizes being slightly underestimated. The use of the 'trim and fill' method, however, showed a significant number of 'missing' studies for 38 % of RE model meta-analyses and 50 % of FE model meta-analyses. On average, the number of studies missing was significantly greater for RE models. Previously, the Begg–Mazumdar correlation has been the main test used to detect possible publication bias. As expected, we found significant agreement between estimates of publication bias based on this correlation and those based on the 'trim and fill' approach. (At present though there is insufficient data to determine whether the Begg–Mazumdar correlation can be used as a 'short-cut' to predict the effect of adding 'missing' studies in terms of the robustness of recalculated effect sizes.) Furthermore, individually significant Begg–Mazumdar correlations also suggest a publication bias in 35 % of the 40 meta-analyses. The asymmetry in funnel graphs previously described for specific topics in ecology and evolution (e.g. Palmer, 2000; Jennions *et al.*, 2001) therefore appears to reflect a more general phenomenon (Jennions & Møller, 2002). Although alternative explanations for a skewed funnel graph should not be neglected (Thornhill *et al.*, 1999), publication bias could be a potential problem for at least one in three meta-analyses.

Correcting for publication bias using the 'trim and fill' method of Duvall & Tweedie (2000*a*, *b*) led to three out of 35 and eight out of 38 initially significant weighted mean effects becoming non-significant for FE and RE models, respectively. The weighted mean effect size was no longer robust after being recalculated for 21 % of FE models and 15 % of RE models. If these findings can be generalized to future studies then 15–21 % of meta-analyses in ecology and evolution could reach erroneous conclusions if no correction is made for publication bias. By contrast, Sutton *et al.* (2000*a*) suggested that only 5–10 % of medical meta-analyses might have reached an incorrect interpretation because of publication bias.

There was one significant and unexpected difference between FE and RE models. For FE models, 71 % of recalculated weighted means were less robust because missing studies reduced the mean effect size. By contrast, for RE models, 59 % were the same or more robust. For RE models, the 'missing studies' were often larger than the initial weighted mean. This is not as predicted by publication bias and suggests that asymmetry in funnel plots may have other, as yet unknown, causes. Sutton *et al.* (2000*a*) only reported decreases in weighted mean effect size in an analysis of 48 medical meta-analyses, although this was not the case when one analyses other datasets (R. Tweedie, personal communication). Chance asymmetry provides an insufficient explanation for our findings because in eight of the 13 cases where missing studies had effect sizes smaller than the weighted mean effect size the number of studies missing was significantly more than expected by chance. The difference between the medical studies of Sutton *et al.* (2000*a*) and the ecological/evolutionary studies analysed here may relate to differences in the range of sample sizes. One or two studies with unusually strong effects and larger sample sizes can generate a weighted mean effect that is larger than most of the reported effect sizes.

For example, the removal of three out of 84 cases with large effects and small variances from Gontard-Danek & Møller (1999) changes the situation from an estimate of 14 missing studies larger than the mean and a recalculated weighted mean of $r = 0.43$, to no missing studies and the original weighted mean of $r = 0.35$. The effect of publication bias in ecology and evolution when using RE models for meta-analysis, at least for the currently available data, is therefore idiosyncratic. In eight cases, the weighted mean became statistically non-significant, but in 13 cases the mean stayed significant and was actually more robust after correcting for publication bias.

In general though, we suggest that any increase in robustness be disregarded. The observed asymmetry in the funnel plots could even be due to selective reporting of studies with effects smaller than the average effect, perhaps because of a prejudice or 'backlash' against a well-established idea or so-called 'bandwagons' (Palmer, 2000; Poulin, 2000). For now, however, we think it is best to be conservative and simply ask whether results remain robust after correcting for funnel-plot asymmetry. Sterne (2000) criticized Sutton *et al.* (2000*a*), saying that the 'trim and fill' method leads to false positive claims of missing studies. This is true, although the criticism can partly be responded to by highlighting how often the number of missing studies is too large to be attributed to chance alone (i.e. $P < 0.05$). Here, this occurred for 38 % of the random-effects models. More generally, Sutton *et al.* (2000*b*) emphasise that the critical aim of 'trim and fill' is not to quantify exactly how many studies are really missing. Rather, it is to test whether results are robust and conclusions unchanged when we correct for possible bias. Here, we show that for the preferred RE model approach, 21 % (eight of 38 meta-analyses) were not robust to potential publication bias. We therefore strongly recommend that authors of meta-analyses routinely include estimates of recalculated effect sizes using 'trim and fill' methods. These methods are easy to apply and, along with fail-safe numbers (Rosenthal, 1991) and Begg–Mazumdar correlations (Begg & Mazumdar, 1994), allow readers to decide for themselves how sensitive a reported significant relationship is to potential bias. Finally, reviewers should also be careful about drawing conclusions about factors that lead to heterogeneity in effect sizes. At present, there is no way of knowing what the effect of 'missing' studies will be on tests of significant variation in effect size among different groups. Significant between-group heterogeneity should therefore be assessed in the light of the number of missing studies. Perhaps 'missing' studies could be conservatively assigned so as to reduce between-group heterogeneity to test whether group differences are still statistically significant.

V. CONCLUSIONS

(1) If asymmetry in a 'funnel plot' of effect size against sample size is due to publication bias, the 'trim and fill' method indicates that 38–50 % of data sets have a significant publication bias as indicated by an excess of 'missing' studies.

(2) The more familiar Begg–Mazumdar correlation also showed that, on average, effect sizes are closer to zero as sample size increases. In general, the findings of the 'trim and fill' method and Begg–Mazumdar correlation were in agreement. 'Trim and fill', however, has the advantage that a 'correction' for possible publication bias can be made by adding 'missing' studies.

(3) In 75 % of the cases analysed using fixed-effects models, the addition of missing studies moved the mean effect closer to zero. For random-effects models, the equivalent figure was 57 %. More importantly, 8–21 % of data sets where the original estimate of the average relationship differed significantly from zero became non-significant.

(4) Stated slightly differently, using the conventional definition of a 'robust' result based on Rosenthal's fail-safe numbers, 15–21 % of effect size estimates were no longer robust after being recalculated to include 'missing' studies.

(5) We conclude that publication bias is a potential problem for reviewers. This is most clearly seen when a quantitative reviewing method like meta-analysis is used, but is equally true for narrative reviews.

(6) Reviewers should always test for publication bias. Aside from established techniques, we specifically recommend the use of 'trim and fill'. It is the only method that allows one to estimate conservatively whether publication bias has influenced the conclusions reached by a reviewer. It is easy to use and does not require expensive software or advanced statistical skills.

VI. ACKNOWLEDGEMENTS

We thank Göran Arnqvist, Michael Brett, Isabelle Côté, Peter Curtis, Mark Forbes, Peter Hamback, Nick Jonsson, Julia Koricheva, Dean McCurdy, Fiorenza Micheli, Iago Mosqueira, Robert Poulin, Howie Riessen, Michael Rosenberg, Gina Schalk, Xianzhong Wang and Peter van

Zandt and others we may have inadvertently omitted for kindly providing unpublished information. J. Shykoff, D. Pope. B. Backwell and J. Christy kindly discussed issues of meta-analysis and the general idea behind performing the present study. M.D.J. thanks the Director of STRI for bridging funding. Papers describing the 'trim and fill' method can be downloaded for free at: *http://www.biostat.umn.edu/~tweedie/documents/tweediecurrentpapers.html*.

VII. REFERENCES

Alatalo, R. V., Mappes, J. & Elgar, M. A. (1997). Heritabilities and paradigm shifts. *Nature* **385**, 402–403.

Arnqvist, G. & Wooster, D. (1995). Meta-analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology and Evolution* **10**, 236–240.

Bauchau, V. (1997). Is there a "file drawer problem" in biological research? *Oikos* **79**, 407–409.

Begg, C. B. (1994). Publication bias. In *The Handbook of Research Synthesis* (eds. H. Cooper and L. V. Hedges), pp. 399–409. Russel Sage Foundation, New York.

Begg, C. B. & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088–1101.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2d ed. L. Erlbaum, Hillsdale, New Jersey.

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

DuMochel, W. & Harris, J. (1997). Comments on Givens, G. H., Smith, D. D. and Tweedie, R. L. (1997) Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science* **12**, 244–245.

Duvall, S. & Tweedie, R. (2000*a*). A non-parametric 'trim and fill' method of assessing publication bias in meta-analysis. *Journal of the American Statistical Association* **95**, 89–98.

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

Egger, M., Davey-Smith, G., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.

Fernandez-Duque, E. & Valeggia, C. (1994). Meta-analysis: a valuable tool in conservation research. *Conservation Biology* **8**, 555–561.

Gleser, L. J. & Olkin, I. (1996). Models for estimating the number of unpublished studies. *Statistics in Medicine* **15**, 2493–2507.

Gontard-Danek, M. C. & Møller, A. P. (1999). The strength of sexual selection: A meta analysis of bird studies. *Behavioral Ecology* **10**, 476–486.

Gurevitch, J. & Hedges, L. V. (1999). Statistical issues in ecological meta-analyses. *Ecology* **80**, 1142–1149.

Jennions, M. D. & Møller, A. P. (2002). Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London* B (in press).

Jennions, M. D., Møller, A. P. & Petrie, M. (2001). Sexually selected traits and adult survival: a meta-analysis. *Quarterly Review of Biology* **76**, 3–36.

Light, R. J. & Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research*. Harvard University Press, Cambridge, Massachusetts.

Macaskill, P., Walter, S. D. & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistical Medicine* **20**, 641–654.

Møller, A. P. & Jennions, M. D. (2001). Testing and adjusting for publication bias. *Trends in Ecology and Evolution* **16**, 580–586.

N.R.C. Committee on Applied and Theoretical Statistics. (1992). *Combining Information: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, D.C.

Palmer, A. R. (1999). Detecting publication bias in meta-analysis: a case study of fluctuating asymmetry and sexual selection. *American Naturalist* **154**, 220–233.

Palmer, A. R. (2000). Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Reviews of Ecology and Systematics* **31**, 441–480.

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? *Proceedings of the Royal Society of London* B **267**, 787–792.

Rosenberg, M. S., Adams, D. C. & Gurevitch, J. (2000). *MetaWin: Statistical Software for Meta-analysis*. Version 2.0. Sinauer Associates, Massachusetts.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research*. Sage Foundation, Newbury Park, California.

Simmons, L. W., Tomkins, J. L., Kotiaho, J. S. & Hunt, J. (1999). Fluctuating paradigm. *Proceedings of the Royal Society of London* B **266**, 593–595.

Song, F., Eastwood, A. J., Gilbody, S., Duley, L. & Sutton, A. J. (2000). Publication and related biases. *Health Technology Assessment* **4** (10), 1–115.

Song, F. & Gilbody, S. (1998). Increase in studies of publication bias coincided with increasing use of meta-analysis. *British Medical Journal* **316**, 471.

Sterne, J. A. C. (2000). High false positive rate for trim and fill method. *British Medical Journal* (Electronic letter at URL: www.bmj.org/cgi/eletters/320/7249/1574).

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R. & Jones, D. R. (2000*a*). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal* **320**, 1574–1577.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R. & Jones, D. R. (2000*b*). High false positive rate for trim and fill method. *British Medical Journal* (Electronic letter at URL: www.bmj.org/cgi/eletters/320/7249/1574).

Thornhill, R., Møller, A. P. & Gangestad, S. (1999). The biological significance of fluctuating asymmetry and sexual selection: A reply to Palmer. *American Naturalist* **154**, 234–241.

Trengenza, T. (1997). Darwin a better name than Wallace? *Nature* **385**, 480.

VIII. APPENDIX: THE DATA SETS

Data sets were taken from the following 40 peer-reviewed meta-analyses. The necessary information was either in the published paper or generously made available to us by the authors.

Arnqvist, G. & Nilsson, T. (2000). The evolution of polyandry: multiple mating and female fitness in insects. *Animal Behaviour* **60**, 145–164.

Arnqvist, G., Rowe, L. & Krupa, J. J. & Sih, A. (1996). Assortative mating by size: a meta-analysis of mating patterns in water striders. *Evolutionary Ecology* **10**, 265–284.

Boissier, J., Morand, S. & Mone, H. (1999). A review of performance and pathogenicity of male and female *Schistosoma mansoni* during the life cycle. *Parasitology* **119**, 447–454.

Brett, M. T. & Goldman, C. (1996). A meta-analysis of the freshwater trophic cascade. *Proceedings of the National Academy of Sciences U.S.A.* **93**, 7723–7726.

Côté, I. M. & Poulin, R. (1995). Parasitism and group size in social animals: a meta-analysis. *Behavioral Ecology* **6**, 159–165.

Côté, I. M. & Sutherland, W. J. (1997). The effectiveness of removing predators to protect bird populations. *Conservation Biology* **11**, 395–405.

Curtis, P. S. (1996). A meta-analysis of leaf gas exchange and nitrogen in trees grown under elevated carbon dioxide. *Plant, Cell and Environment* **19**, 127–137. (Using the data file at URL: http://cdiac.esd.ornl.gov/epubs/ndp/ndp072/ndp072.html)

Curtis, P. S. & Wang, X. (1998). A meta-analysis of elevated CO_2 effects on woody plant mass, form, and physiology. *Oecologia* **113**, 299–313. (Using the data file at URL: http://cdiac.esd.ornl.gov/epubs/ndp/ndp072/ndp072.html)

Fiske, P., Rintamäki, P. & Karvonen, E. (1998). Mating success in lekking males: a meta-analysis. *Behavioral Ecology* **9**, 328–338.

Gontard-Danek, M. C. & Møller, A. P. (1999). The strength of sexual selection: a meta-analysis of bird studies. *Behavioral Ecology* **10**, 476–486.

Gurevitch, J. & Hedges, L. V. (1993). Meta-analysis: combining the results of independent experiments. In *Design and Analysis of Experiments* (ed. by S. Scheiner and J. Gurevitch), pp. 378–398. Chapman and Hall, New York. (Data as presented in Rosenberg *et al*., 2000.)

Harper, D. G. C. (2000). Feather mites, pectoral muscle condition, wing length and plumage coloration of passerines. *Animal Behaviour* **58**, 553–562.

Järvinen, A. (1991). A meta-analytic study of the effects of female age on laying-date and clutch-size in the Great Tit *Parus major* and the Pied Flycatcher *Ficedula hypoleuca*. *Ibis* **133**, 62–67.

Jennions, M. D., Møller, A. P. & Petrie, M. (2001). Sexually selected traits and adult survival: a meta-analysis of the phenotypic relationship. *Quarterly Review of Biology* **76**, 3–36.

Koricheva, J. (2002). Meta-analysis of sources of variation in fitness costs of plant antiherbivore defenses. *Ecology* **83**, 176–190.

Koricheva, J., Larsson, S. & Haukioja, E. (1998). Insect performance on experimentally stressed woody plants: a meta-analysis. *Annual Review of Entomology* **43**, 195–216.

Koricheva, J., Larsson, S., Haukioja, E. & Keinanen, M. (1999). Regulation of woody plant secondary metabolism by resource availability: Hypothesis testing by means of meta-analysis. *Oikos* **83**, 212–226.

Leung, B. & Forbes, M. R. (1996). Fluctuating asymmetry in relation to stress and fitness: effects of trait type as revealed by meta-analysis. *Ecoscience* **3**, 400–413. (Using the data file at URL: www.biology.ualberta.ca/palmer.hp/DataFiles.htm)

Micheli, F. (1997). Eutrophication, fisheries and consumer-dynamics in marine pelagic ecosystems. *Science* **285**, 1396–1398.

Møller, A. P. (1999). Asymmetry as a predictor of growth, fecundity and survival. *Ecology Letters* **2**, 149–156.

Møller, A. P. (2000). Developmental stability and pollination. *Oecologia* **123**, 149–157.

Møller, A. P. & Alatalo, R. V. (1999). Good genes effects in sexual selection. *Proceedings of the Royal Society of London B* **266**, 85–91.

Møller, A. P., Christe, P., Erritzøe, J. & Mavarez, J. (1998). Condition, disease and immune defence. *Oikos* **83**, 301–306.

Møller, A. P., Christe, P. & Lux, E. (1999). Parasitism, host immune function, and sexual selection. *Quarterly Review of Biology* **74**, 3–20.

Møller, A. P. & Ninni, P. (1998). Sperm competition and sexual selection: a meta-analysis of paternity studies of birds. *Behavioural Ecology and Sociobiology* **43**, 345–358.

Møller, A. P. & Shykoff, J. A. (1999). Morphological developmental stability in plants: patterns and causes. *International Journal of Plant Sciences* **160**, S135–S146.

Møller, A. P. & Thornhill, R. (1998). Bilateral symmetry and sexual selection: a meta-analysis. *American Naturalist* **151**, 174–192.

Mosqueira, I. Côté, I. M., Jennings, S. & Reynolds, J. D. (2000). Conservation benefits of marine reserves for fish populations. *Animal Conservation* **3**, 321–332.

Osenberg, C. W., Sarnelle, O., Cooper, S. D. & Holt, R. D. (1999). Resolving ecological questions through meta-analysis: goals, metrics, and models. *Ecology* **80**, 1105–1117. (Data set 1.)

Osenberg, C. W., Sarnelle, O., Cooper, S. D. & Holt, R. D. (1999). Resolving ecological questions through meta-analysis: goals, metrics, and models. *Ecology* **80**, 1105–1117. (Data set 2.)

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? *Proceedings of the Royal Society Biological Sciences Series B* **267**, 787–792.

Poulin, R. (2000). Variation in the intraspecific relationship between fish length and intensity of parasitic infection: biological and statistical causes. *Journal of Fish Biology* **56**, 123–137.

Riessen, H. P. (1999). Predator-induced life history shifts in *Daphnia*: a synthesis of studies using meta-analysis. *Canadian Journal of Fisheries and Aquatic Sciences* **56**, 2487–2494.

Schmitz, O. J., Hamback, P. A. & Beckerman, A. P. (2000). Trophic cascades in terrestrial systems: a review of the effects of carnivore removals on plants. *American Naturalist* **155**, 141–153.

Sokolovska, N., Rowe, L. & Johansson, F. (2000). Fitness and body size in mature odonates. *Ecological Entomology* **25**, 239–248.

Thornhill, R. & Møller, A. P. (1999). The relative importance of size and asymmetry in sexual selection. *Behavioral Ecology* **9**, 546–551. (Only effect sizes for size.)

Van der Werf, E. (1992). Lack's clutch size hypothesis: an examination of the evidence using meta-analysis. *Ecology* **73**, 1699–1705.

Van Zandt, P. A. & Mopper, S. (1998). A meta-analysis of adaptive deme formation in phytophagous insect populations. *American Naturalist* **152**, 595–604.

Vøllestad, L. A., Hindar, K. & Møller, A. P. (1999). A meta analysis of fluctuating asymmetry in relation to heterozygosity. *Heredity* **83**, 206–218.

Wang, X. & Curtis, P. S. (2002). A meta-analytical test of elevated CO_2 effects on plant respiration. *Plant Ecology* (in press).