A survey of publication bias within evolutionary ecology

Phillip Cassey^{1*}, John G. Ewen²,
Tim M. Blackburn¹ and Anders P. Møller²

¹*School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

²*Laboratoire d'Ecologie Evolutive Parasitaire, Université Pierre et Marie Curie, Paris, France*

* Author for correspondence (cassey@votan.ens.fr).

Recd 28.04.04; Accptd 17.05.04; Published online 15.07.04

Publication bias has been recognized as a problem in ecology and evolution that can undermine reviews of research results. Unfortunately, direct tests of publication bias are extremely rare. Here, we quantify a well-discussed but, to our knowledge, previously untested form of publication bias: the publication of results with and without estimates of effect size. We find that results published without effect sizes are a biased sample of those that are published. This further complicates the already difficult task of compiling quantitative literature reviews and meta-analytic studies.

Keywords: meta-analysis; effect size; reporting bias; within-study publication bias

1. INTRODUCTION

Combining research results is of fundamental importance in ecology as most studies are neither large enough nor cover a sufficiently wide range of conditions to provide ecological generalizations (Gates 2002). In particular, the development of meta-analytical techniques has improved our ability to synthesize often scattered and heterogeneous research findings. In ecology and evolutionary biology, meta-analyses have provided a strong quantitative framework for testing the statistical significance of combined results. Empirical examples of meta-analyses have become considerably more numerous since their introduction in the early 1990s (Arnqvist & Wooster 1995) and now number over 100 (Gurevitch *et al.* 2001; Gates 2002). Recently, several critical reviews have been published that assist ecologists in conducting quantitative summaries of the literature and adopting meta-analytic approaches (Gurevitch & Hedges 1999; Osenberg *et al.* 1999; Gurevitch *et al.* 2001; Møller & Jennions 2001; Gates 2002). Importantly, the synthesis of published results relies on the presentation of adequate information in previous studies to allow the results to be quantified.

Ecologists have traditionally relied on *p*-values for interpreting the significance of null hypothesis tests (and applied data analysis has been predicted to remain, for the most part, solidly frequentist for the foreseeable future (Hoenig & Heisey 2001)). For example, Anderson *et al.* (2000) estimated that an average of over 6000 *p*-values appeared in every annual volume of the journal *Ecology* between 1988 and 1997. The small to moderate sample

sizes (*n*) and effect sizes (δ) that are usually reported in ecological studies result in a situation where many, if not most researchers, will fail to reject their null hypotheses at $\alpha = 0.05$. In a study of the amount of variance that main effects generally explained in ecological studies, Møller & Jennions (2002) found that the average sample size that is required to detect a statistically significant effect is commonly larger than is considered in many, perhaps most, ecological studies. It might seem intuitive that the outcomes of standard statistical significance tests in individual studies could be used to assess the consistency of effect magnitudes by sorting the studies into categories according to the significance of the tests. However, the *p*-value does not provide information about either the size or the precision of the estimated effect, and the dichotomy that is sometimes perceived to exist between 'significant' and 'non-significant' is false (Johnson 1999). Similarly, the logic of comparing multiple significance values (vote-counting) is flawed when *n* is small, and is always wasteful of information (Hedges & Olkin 1985). Notably, conclusions based on the consistency of *p*-value magnitudes can be highly misleading (see Humphreys (1980) for a more detailed discussion of this fallacy and Osenberg *et al.* (1999) for an ecological example). Importantly, there is no easy way to tell whether results of studies are consistent with one another from the outcomes of their individual significance tests alone.

The objective of any quantitative review is to summarize estimates of the standardized magnitude of an ecological response (i.e. the 'effect size') relative to a given correlation or manipulation variable. The effect size, e_i from each of the $i = 1, \dots, k$ study variables included, is simply the magnitude of the change in the response re-expressed to remove the arbitrary scale dependence. There are no pre-conditions on the definition of e_i , and it has even been suggested that, given the diversity of ecological hypotheses, there is little reason to limit the diversity of metrics of effect size (Osenberg *et al.* 1999). In meta-analyses, the standardized effect size is the response that is subjected to analysis for the comparison of results from a given research question (Hedges & Olkin 1985).

There is a problem, however, when researchers publish results that provide limited or no information regarding the statistics that the readers and reviewers require for comparing results. It has recently been observed that many authors are failing to present results with sufficient information for others to calculate effect sizes (Anderson *et al.* 2001; Colegrave & Ruxton 2003). This has the potential to generate 'publication bias', a term that has traditionally been used to indicate when the strength or direction of the results of published and unpublished studies differs (Song *et al.* 2000; Møller & Jennions 2001). However, another form of publication bias may result if a difference exists between results that are published with and without effect size information. Results published without an effect size, or the sign of the effect, or a measure of precision, are referred to as 'naked *p*-values', and are the worst example of biased result reporting in frequentist studies (Anderson *et al.* 2001). To facilitate the synthesis of ecological research questions, presentation of results in this manner must be avoided at all times (Colegrave & Ruxton 2003). Nevertheless, naked *p*-values are still published and the magnitude of the publication bias that may accrue from differences between their effect sizes and those attached to properly reported results is so

far unquantified. Here, we present a unique example of the magnitude of publication bias that can manifest within the primary ecological literature.

2. AN EXAMPLE FROM EVOLUTIONARY ECOLOGY

We collated primary studies of facultative sex-ratio adjustment in birds that were published since the development of molecular sexing techniques, from searches of the 'ISI' and 'current contents' databases. We followed this with a thorough examination of the bibliographies of the studies, and correspondence with their authors. Our aim was to provide a quantitative review of sex-ratio adjustment by female birds and assess the role of potential ecological moderator variables (Ewen *et al.* 2004). The response variable (P_m) in these studies is the proportion of one sex (i.e. male) that a female produces in a given clutch. The hypothetical relationship between this response and a series of potential moderator variables (as many as 18 in a single study; see Lessells *et al.* 1996) is then assessed. Because the individual observations (i.e. nestlings within broods) are assumed to be distributed binomially, multiple logistic regression-type approaches (χ^2 - and F -test test statistics) are the most common method of analysis (e.g. Wilson & Hardy 2002; Ewen *et al.* 2003). In total, we found 52 studies that had reported the results of relationships between sex-ratio variability and predicted independent variables since the advent of molecular sexing methods.

Out of the 52 studies that we found, only 25 (48%) presented enough information to calculate direct estimates of effect sizes for all of their independent variables. On average, sex-ratio variability was analysed against more than six univariate variables per study. Out of these variables, only seven (2%) were presented without means, differences, effect sizes, test statistics or precision (i.e. naked p -values). However, out of the 293 variables for which some information was available, 153 (52%) were presented without either estimates or notation for the trend of the effect. Therefore, on initial examination of the primary literature, effect-size estimates could be reasonably calculated for only 47% of the original statistical hypotheses.

Research summaries are based on the assumption that the literature reviewed is unbiased (Rosenthal 1991). However, as noted in § 1, ecological reviews can suffer substantially from the publication bias that occurs whenever the strength or direction of the results of published and unpublished studies differ (Møller & Jennions 2001; Kotiaho & Tomkins 2002). The so-called 'file-drawer' problem arises when studies remain unpublished because decisions at the submission, editorial or referee stage lead to a scientific literature that is a biased sample of the available studies (Csada *et al.* 1996; Bauchau 1997). Consequently, there are several methods available to the researcher for testing for publication bias, and it has been strongly recommended that if bias is detected, further analysis and interpretation only be undertaken with extreme caution (Kotiaho & Tomkins 2002).

For the relationship between facultative sex-ratio adjustment and independent variables, Ewen *et al.* (2004) considered only studies that could confirm that they had measured primary sex ratio ($n = 40$). We conducted the

rank correlation test of Begg & Mazumdar (1994) to investigate the association between standardized effect size and sample size for all the variables for which effect sizes could be directly calculated from the original sources. The correlation was highly significant (figure 1a; $r = -0.32$, $n = 79$, $p < 0.01$), indicating that there are fewer than the expected number of studies with negative effects at low sample sizes, and hence a publication bias.

Ewen *et al.* (2004) then supplemented their data by email communication with the corresponding authors from all of the studies of primary sex-ratio adjustment that did not present enough published information to calculate effect sizes. When the additional effect-size estimates from this correspondence were included in the analysis, there was no association between standardized effect size and increasing sample size (figure 1b; $r = -0.06$, $n = 214$, $p > 0.4$). The proportion of missing effect-size estimates ranged from 0.0 to 1.0 per study, with a mean of 0.36. There was also a significant positive correlation between the proportion of missing estimates and the number of independent variables per study ($r = 0.54$, $n = 40$). Thus, the more variables in a study, the more likely that researchers did not present adequate information to calculate effect-size estimates. Following Koricheva (2003), we also examined the proportion of missing estimates that were either statistically 'significant' or 'non-significant' (univariate $\alpha = 0.05$, without correction for multiple comparisons). We used a generalized linear mixed model (SAS v. 8.02; Littell *et al.* 1996) to model the proportion of missing estimates (a binary variable; 0 = missing, 1 = present) that were either significant or non-significant, including each study as a random variable. We found that, within published manuscripts, researchers were considerably more likely not to publish effect-size estimates if their results were non-significant (estimate = 4.21, s.e. = 0.62, $p < 0.01$). By contrast, Koricheva (2003) found no evidence for publication bias against non-significant results among published and non-published dissertation manuscripts. If both of these results are found to be common within the evolutionary ecology literature then publication bias is currently more of a problem within journals than within 'file-drawers'.

3. CONCLUSION

Previously, direct tests of publication bias have been rare (but see Møller & Thornhill 1998; Jennions *et al.* 2001; Koricheva 2003) because of the difficulty involved in obtaining missing (i.e. unpublished) results. However, publication bias is not just a 'file-drawer' problem but can also be manifest within the primary literature. This is particularly likely to be a problem in research fields where results are presented for a large number of independent test variables, such as in ecology (e.g. Anderson *et al.* 2001). By directly quantifying a sample of missing effect-size estimates we have shown that they are a biased sample of published results. If general, this problem further contributes to the already difficult issue of publication bias in ecological studies (Kotiaho & Tomkins 2002). Although none of the statistical concepts we discuss is novel, the degree of bias in our example is clearly arresting. In the past, researchers have relied on contacting authors for missing information (e.g. Ewen *et al.* 2004). Nevertheless,

Figure 1. Funnel plots of the relationship between sex-ratio adjustment and independent traits for (a) effect sizes that could be directly calculated from the original sources, and (b) with the additional effect-size estimates from email correspondence with the authors who did not present enough published information to calculate effect sizes. Positive values reflect male-biased sex ratios and negative values reflect female-biased sex ratios. Overlaid significance lines are calculated for $\alpha = 0.05$ following Sutton (1990).

even outside any meta-analysis, the reader must be able adequately to evaluate the biological significance of an effect. We reiterate that the presentation of informative test statistics and their associated confidence intervals is particularly important if authors and editors are to facilitate the quantitative interpretation and synthesis of research results.

Acknowledgements

The authors thank Pablo Inchausti, Andy Gonzalez, and Gerard Lacroix for discussions contributing to the ideas in this commentary. We thank two anonymous reviewers for their comments. P.C. was funded by a CNRS postdoctorate with the group 'Eco-Evolution Mathématique' and the Leverhulme Trust (grant no. F/00094/AA). P.C. gratefully acknowledges the support of Rebecca Boulton.

Anderson, D. R., Burnham, K. P. & Thompson, W. L. 2000 Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Mgmt* **64**, 912–923.

Anderson, D. R., Link, W. A., Johnson, D. H. & Burnham, K. P. 2001 Suggestions for presenting the results of data analysis. *J. Wildl. Mgmt* **65**, 373–378.

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

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

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

Colegrave, N. & Ruxton, G. D. 2003 Confidence intervals are a more useful complement to non-significant tests than are power calculations. *Behav. Ecol.* **14**, 446–450.

Csada, R., 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.

Ewen, J. G., Crozier, R. H., Cassey, P., Ward-Smith, T., Painter, J. N., Robertson, R. J., Jones, D. A. & Clarke, M. F. 2003 Facultative control of offspring sex in the cooperatively breeding bell miner, *Manorina melanophrys*. *Behav. Ecol.* **14**, 157–164.

Ewen, J. G., Cassey, P. & Møller, A. P. 2004 Facultative primary sex ratio variation: a lack of evidence in birds? *Proc. R. Soc. Lond. B* **271**, 1277–1282. (DOI 10.1098/rspb.2004.2735.)

Gates, S. 2002 Review of methodology of quantitative reviews using meta-analysis in ecology. *J. Anim. Ecol.* **71**, 547–557.

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

Gurevitch, J., Curtis, P. S. & Jones, M. H. 2001 Meta-analysis in ecology. *Adv. Ecol. Res.* **32**, 199–247.

Hedges, L. V. & Olkin, I. 1985 *Statistical methods for meta-analysis*. San Diego, CA: Academic.

Hoenig, J. M. & Heisey, D. M. 2001 The abuse of power: the pervasive fallacy of power calculation for data analysis. *Am. Statistician* **55**, 19–24.

- Humphreys, L. G. 1980 The statistics of failure to replicate: a comment on Buriel's (1978) conclusions. *J. Educat. Psychol.* **72**, 71–75.
- Jennions, M. D., Møller, A. P. & Petrie, M. 2001 Sexually selected traits and adult survival: a meta-analysis of the phenotypic relationship. *Q. Rev. Biol.* **76**, 3–36.
- Johnson, D. H. 1999 The insignificance of statistical significance testing. *J. Wildl. Mngmt* **63**, 763–772.
- Koricheva, J. 2003 Non-significant results in ecology: a burden or a blessing in disguise? *Oikos* **102**, 397–401.
- Kotiaho, J. S. & Tomkins, J. L. 2002 Meta-analysis, can it ever fail? *Oikos* **96**, 551–553.
- Lessells, C. M., Mateman, A. C. & Visser, J. 1996 Great tit hatchling sex ratios. *J. Avian Biol.* **27**, 135–142.
- Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. 1996 *SAS system for mixed models*. Cary, NC: SAS Institute.
- Møller, A. P. & Jennions, M. D. 2001 Testing and adjusting for publication bias. *Trends Ecol. Evol.* **16**, 580–586.
- Møller, A. P. & Jennions, M. D. 2002 How much variance can be explained by ecologists and evolutionary biologists? *Oecologia* **132**, 492–500.
- Møller, A. P. & Thornhill, R. 1998 Bilateral symmetry and sexual selection: a meta-analysis. *Am. Nat.* **151**, 174–192.
- 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.
- Rosenthal, R. 1991 *Meta-analytic procedures for social research*. New York: Russell Sage Foundation.
- Song, F., Eastwood, A., Gilbody, S., Duley, L. & Sutton, A. 2000 Publication and related biases. *Hlth Technol. Assess.* **4**, 1–115.
- Sutton, J. B. 1990 Values of the index of determination at the 5% significance level. *Statistician* **39**, 461–463.
- Wilson, K. & Hardy, I. C. W. 2002 *Statistical analysis of sex ratios: an introduction in sex ratios: concepts and research methods*. Cambridge University Press.