Notes and Comments

Detecting Publication Bias in Meta-analyses: A Case Study of Fluctuating Asymmetry and Sexual Selection

A. Richard Palmer[1,2,*]

1. Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada;
2. Bamfield Marine Station, Bamfield, British Columbia V0R 1B0, Canada

Submitted April 15, 1998; Accepted October 29, 1998

Keywords: developmental stability, mating behavior, sexual signaling, selective reporting, funnel graph, investigator effects, statistical methods.

Those familiar with human nature and the publication process acknowledge that biases due to selective reporting of results are likely widespread in all fields of academic inquiry that depend on tools of statistical inference (e.g., see Begg and Berlin 1988; Iyengar and Greenhouse 1988; and the extensive discussion following each). However, although qualitative and quantitative methods exist for assessing the prevalence of selective reporting, and selective reporting has been studied in the medical and social sciences (Begg 1994, and references therein), the issue has received little attention in recent meta-analyses of ecological and evolutionary patterns. Clearly, "if publication bias is present, and if it operates in the same direction for all studies (as is likely), then [meta-analysis] is likely not only to produce biased summary estimates but also to produce estimates which are apparently precise and accurate leading to conclusions which may not only be wrong but appear convincing" (Begg and Berlin 1988, p. 437).

To biologists unacquainted with the formal study of publication patterns, the terms "selective reporting" (statistical significance of an outcome influences its likelihood of being reported or published) and "publication bias" (the inflation of average effect size due to selective reporting) may imply a conscious intent to deceive, but this is unwarranted. The phenomenon of selective reporting, and the biases that may ensue, need not be the fault of individual investigators (Begg and Berlin 1988; Iyengar and Greenhouse 1988; Begg 1994). Both reviewers and editors may be hesitant to accept nonsignificant results for publication, particularly when they are based on small sample sizes. Also, authors are, understandably, more likely to submit results based on small sample sizes if they are significant statistically than if they are not (the "file drawer" problem; Rosenthal 1979). Therefore, most investigators are likely guilty of selective reporting to some extent. The crucial questions for meta-analysis are: to what extent does selective reporting bias estimates of average effect size and artificially inflate estimates of effect size heterogeneity?

Meta-analysis also permits the impact of particular authors on a field in which certain authors have contributed a high proportion of the results. Here, too, however, caution must be exercised when interpreting differences: different effect sizes from different authors need not imply any dubious conduct. Different investigators legitimately use different methods in different systems, and some may be better than others at choosing methods or systems of study that are more likely to yield a strong signal.

Useful as it may be, meta-analysis is not without its limitations. Strictly speaking, quantitative meta-analytic estimates of mean effect size by themselves (e.g., "The average correlation between body size and fitness is 0.44.") may not be very meaningful for most questions in ecology and evolution (Leamy 1997). The statistical models on which meta-analyses are based assume that a single underlying "true" effect size exists and that estimates of effect size exhibit "random" variation among studies due to differences in rigor of sampling design, data collection, and analysis, as well as to unavoidable sampling error (Rosenthal 1991). Clearly, the true correlation, for example, between body size and fitness is likely to differ significantly among taxa and traits. So a single average estimate of a truly heterogeneous phenomenon, whether statistically significant or not, is of little predictive value.

Meta-analytic methods do, however, permit quantitative

* E-mail: Rich.Palmer@UAlberta.CA.

Am. Nat. 1999. Vol. 154, pp. 220–233.

tests of heterogeneity (Rosenthal 1991); for example, "Do the differences between males and females in the dependence of fitness on body size contribute significantly to the heterogeneity of effect size?" But here too, heterogeneity in effect sizes due to variation among taxa, traits, or study conditions may obscure differences between the sexes, so conclusions about the absence of differences may not be very robust. Nonetheless, meta-analysis offers a substantial improvement over nonquantitative narrative summaries of the literature because it obliges authors of reviews to present results in standardized units and to be explicit about sources of data, and methods of weighting and analysis (Cooper and Hedges 1994; Arnqvist and Wooster 1995).

Approaches to the Problem of Selective Reporting

Because estimates of mean effect size and of effect size heterogeneity may be compromised by biases due to selective reporting, such bias must be ruled out or minimized before drawing conclusions about effect size variation. Vevea and Hedges (1995, p. 420) identify three classes of methods that address the problem: those for detecting selective reporting, those that attempt to eliminate the effect of selective reporting, and those that compensate for the impact of selective reporting (i.e., compensate for bias).

Detecting Selective Reporting

Light and Pillemer (1984) introduced a particularly attractive qualitative approach for detecting selective reporting: the "funnel graph." It has become a highly recommended component of preliminary exploratory analyses that should precede a formal meta-analysis (Begg 1994).

In the absence of selective reporting, familiar statistical principles offer three simple and straightforward predictions (see fig. 1*A*): as sample size decreases, the variation about the "true" effect size should increase owing to increased sampling error; average effect size should be independent of sample size; and regardless of sample size, individual effect sizes should exhibit a normal distribution about the "true" mean effect size due to random sampling error. In addition, because results in a meta-analysis are transformed to a standardized statistic, such as the Pearson product-moment correlation coefficient (r), tabled critical values provide a convenient reference point against which to view effect sizes and assess the extent of selective reporting (*curved dashed lines*, figs. 1–4).

Selective reporting induces two departures from the statistically unbiased pattern (fig. 1*A*), depending on the "true" underlying effect size (Light and Pillemer 1984). First, if the true effect size is weak (mean $r < 0.1$; Arnqvist and Wooster 1995), then selective reporting yields fewer than expected nonsignificant results (those lying inside the 95% significance thresholds; fig. 1*B*). Underreporting should be most pronounced at small sample sizes because authors will have less confidence in a nonsignificant result at small sample size and be less likely to submit them for publication, and because reviewers and editors will be less inclined to accept such weakly supported nonsignificant results. Therefore, when the "true" effect size is small, selective reporting yields statistically significant departures of effect sizes from normality at small sample sizes (fig. 1*B*).

Second, if the true effect size is moderate ($r \geq 0.25$; Arnqvist and Wooster 1995), extreme effect size values that had previously reached statistical significance on either side of 0 when the "true" effect size was small (fig. 1*B*) no longer reach statistical significance on the side of 0 opposite the "true" effect size and are thus less likely to be reported (lower portion of scatter in fig. 1*C*). In other words, although the distribution of actual effect sizes remains symmetrical about the "true" mean effect size, this distribution is now shifted relative to the 95% significance thresholds that remain symmetrical around 0. So, unlike the case for a weak "true" effect size (fig. 1*B*), selective reporting now yields a statistically significant dependence of effect size on sample size (fig. 1*C*).

As should be evident from figure 1*B*, the absence of a dependence of effect size on sample size may not mean selective reporting is absent. Ironically, statistical evidence of a dependence of effect size on sample size (fig. 1*C*) suggests that the "true" effect size is likely modest, though it is hardly strong support for such a conclusion.

This funnel graph approach of Light and Pillemer (1984) and its quantitative extensions—tests for departures from normality as a function of sample size and tests for dependence of mean effect size on sample size (Begg 1994)—should really be used more widely to reduce the likelihood of being misled by tabulated statistical summaries in meta-analyses of ecological and evolutionary patterns (e.g., see Arnqvist et al. 1996).

Eliminating or Correcting for Selective Reporting

The effect of selective reporting may potentially be eliminated by incorporating as many unpublished results as possible in a meta-analysis (Begg 1994). This approach is preferred because it provides the most reliable estimate of true effect size. It may be particularly valuable in medical research where registries of funded studies exist (Cooper and Hedges 1994). However, because many "exploratory" studies seem likely to be conducted as a routine part of research in ecology and evolution, an exhaustive search for unpublished studies would seem impractical.

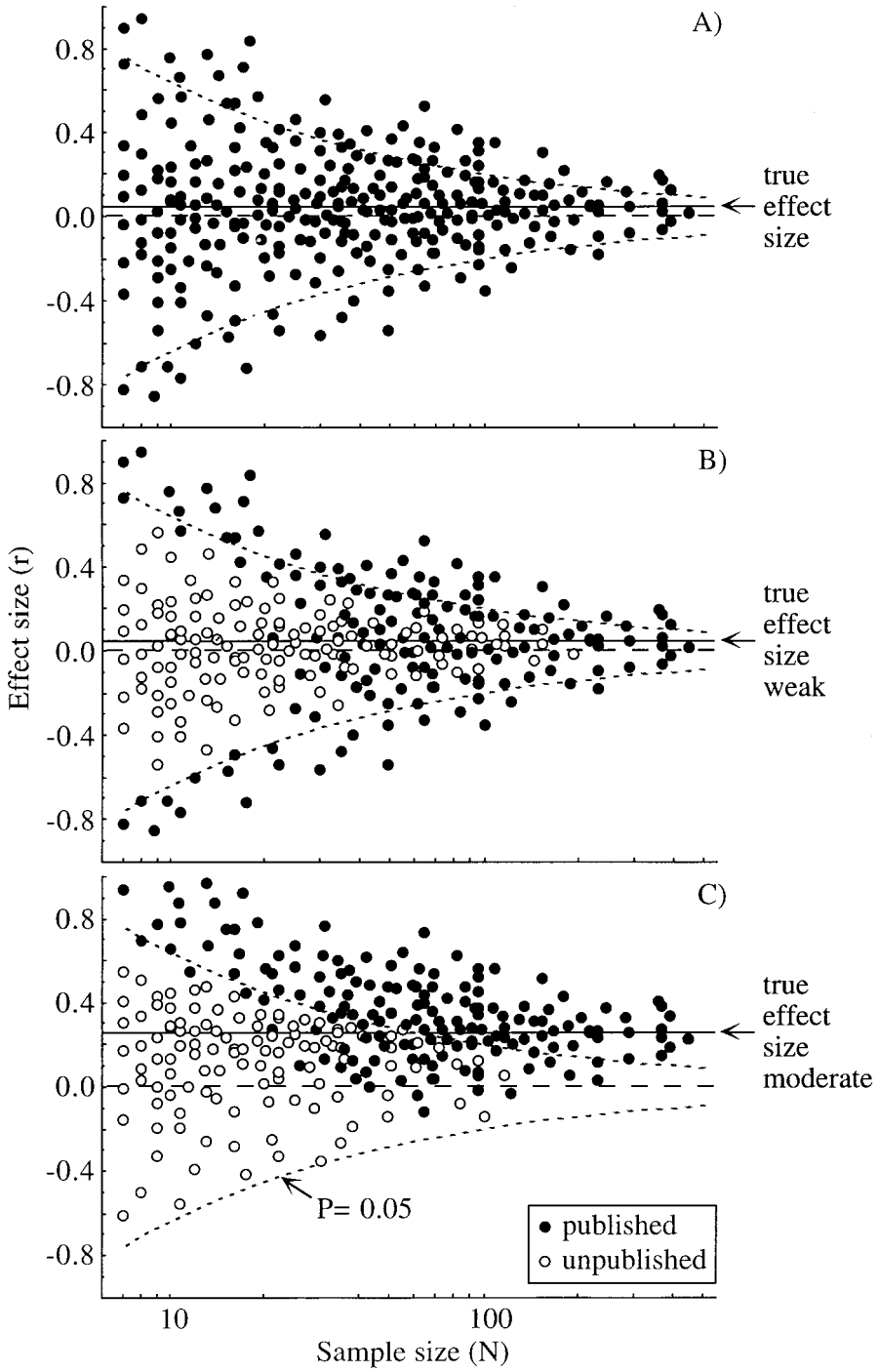

Figure 1: Hypothetical funnel graphs (modified from Light and Pillemer 1984): the distribution of effect size (*r*) as a function of sample size (*N*) for three situations. *A*, "True" effect size weak, selective reporting absent; *B*, "true" effect size weak, selective reporting present (the classical funnel pattern); and *C*, "true" effect size moderate, selective reporting present (one side of funnel missing). Selective reporting refers to a reduced likelihood of publication if effect size is not significant statistically (*open circles*). Curves for 95% significance thresholds were constructed from table 25 of Rohlf and Sokal (1981).

More elaborate analytic methods that attempt to correct for the effects of selective reporting before computing mean effect sizes, and that do not depend on the uncovering of unpublished studies, offer an alternative approach (e.g., Vevea and Hedges 1995). Most promising among these is one based on weighted distribution theory (Begg 1994), in which the true distribution of effect sizes is estimated by assuming that the likelihood of publication depends on, for example, the statistical significance of an outcome: the lower the statistical significance, the lower the probability of publication. Different weight functions to describe an author's desire to publish may be specified, such as a binary weight for studies above or below some critical threshold, or a continuously varying weight that declines linearly or exponentially as a function of the statistical significance of an outcome. Unfortunately, these methods are "for the most part … very new, and their statistical properties have not been subjected to rigorous scrutiny" (Begg 1994, p. 405). In addition, their validity depends critically on an accurate knowledge of the form of the weighting function. Therefore, quantitative corrections for publication bias should probably be applied with caution.

Interpreting Results in the Presence of Selective Reporting

The traditional meta-analytic test for a "fail-safe" number of studies (Cooper 1979; Rosenthal 1991) provides some help with the interpretation of mean effect size in the presence of possible bias. This widely used technique (Begg 1994) estimates the number of studies having zero effect that would have to be published to reduce the overall mean effect size to nonsignificance. But the fail-safe number, although potentially reassuring, can be misleading for two reasons. First, unreported studies are unlikely all to be of zero effect, so although practical and easy to interpret, the fail-safe number overestimates the number of unreported studies needed to have been put in the "file drawer" (Rosenthal 1979). It should therefore not be interpreted literally as the total number of unreported studies required to eliminate the statistical significance of a mean effect, since the effect sizes of some unreported studies could be of opposite sign and greater magnitude (*open circles*, fig. 1*C*). Second, selective reporting may cause the shape of the frequency distribution of effect sizes to change with sample size (*solid circles*, fig. 1*B, C*). Therefore, pooled Z scores across all studies may not be normally distributed, which will reduce the confidence in a statistic that assumes normality.

Although estimates of the fail-safe number of studies provide at least some peace of mind when interpreting the magnitude of mean effect sizes, qualitative graphic approaches seem considerably more informative and convincing (see "Impact on Estimates of Effect Size Heterogeneity" in the "Discussion"). A case study will help to illustrate this point.

A Case Study of Publication Bias

Studies of fluctuating asymmetry (FA)—subtle deviations from symmetry thought to reflect instability of development (Palmer 1996, and references therein)—now exist in sufficient number that meta-analytic methods can test the significance of overall "effect" sizes for various phenomena of interest, including correlations between FA and sexual selection (Møller and Thornhill 1998), FA and stress or fitness (Leung and Forbes 1996; Møller 1997), and heritability of developmental stability (Møller and Thornhill 1997). However, the deviations from symmetry that give rise to FA are so small (often 1% of trait size or less; Palmer 1996) that concerns about the reported magnitude of associations with FA (Houle 1997) seem legitimate. If many tests were done for associations with asymmetry, but those yielding statistical significance ($P < .05$) were more likely to be published, how valid are general conclusions about associations with asymmetry?

The recent extensive meta-analysis of relations between asymmetry and sexual selection by Møller and Thornhill (1998) offers an opportunity to apply some simple tests for selective reporting and, as a consequence, to assess its extent among studies of the relation between asymmetry and sexual selection. Although other meta-analyses have examined patterns of variation in FA (Leung and Forbes 1996; Møller 1997; Møller and Swaddle 1997; Møller and Thornhill 1997), I restricted my analysis to the most recent one by Møller and Thornhill (1998). Because many samples (146 total) were presumably all tabulated using the same criteria, effect size distributions could be examined from several perspectives to test for selective reporting without the potentially confounding effects of different synthesis protocols in different meta-analyses and different underlying effect size distributions.

Methods

Data Inclusion and Coding

The effect size (r), sample size (N), and grouping variables for all 146 samples were entered as in table 1 of Møller and Thornhill (1998). I did not attempt to verify the conversion of results in all of the original studies to the effect sizes tabulated, but I have assumed that Møller and Thornhill did so in an objective and consistent manner and entered them correctly in the table. To these entries I added two additional grouping variables: inclusion/publication status, and author of study.

To determine inclusion or publication status, samples were identified as excluded, published, or unpublished. Møller and Thornhill excluded six samples from their analyses, but the grounds for exclusion did not seem well founded for four of these (see "Excluded Samples" later). However, to permit direct comparisons with their results, I also excluded these six samples from the analyses presented here. I did not attempt to verify whether samples listed as unpublished were subsequently published.

The author of study variable was used to assess investigator effects. Samples were distinguished by author where more than 10 samples could be attributed to a given author: Markow as coauthor ($N = 15$), Møller as coauthor ($N = 24$), Thornhill as coauthor ($N = 25$). All remaining samples were codes as "other author" ($N = 82$).

Statistical Analyses

To avoid possible spurious results among multiple post hoc analyses, I limited my quantitative analyses to three questions not examined by Møller and Thornhill (1998). First, does the frequency distribution of effect size differ from that expected in the absence of selective reporting? Second, does effect size vary as a function of sample size (e.g., as in fig. 1*C*)? Finally, does the dependence of effect size on sample size differ between traits where correlations with asymmetry were expected by the authors to be strong versus traits where such correlations were expected to be weak or absent?

Observed frequencies of Z_r (Fisher's Z transformation of effect size correlation, r, which is not normally distributed if the parametric mean r is not 0; Zar 1984, p. 239) were compared to a normal distribution whose mean and standard deviation were determined from the data (intrinsic hypothesis), using the Kolmogorov-Smirnov test (K-S test; Sokal and Rohlf 1995, p. 712). Tests against an intrinsic, rather than extrinsic, hypothesis were used so that tests for distribution shape were not confounded if the true mean or SD varied with sample size (see fig. 1*B*, *C*). To determine whether departures from normality depended on sample size, all 140 included samples were ranked by sample size from smallest ($N = 7$) to largest ($N = 500$). A K-S test was conducted on overlapping groups of approximately 20 samples taken in order from smallest to largest N in intervals of 10 (e.g., samples 1–20, 11–30, 21–40, etc.). The mean N of each group was also computed to allow results to be displayed graphically. In the event of ties, adjacent samples were either included in or excluded from a group so as to keep the group as close to 20 samples as possible. Among the 13 groups (see fig. 2*B*), only four deviated from 20 samples by more than one (three included 18 samples, three included 19 samples, six included 20 samples, and one included 26 samples).

The dependence of effect size on sample size was assessed via two nonparametric tests of association recommended by Begg (1994): Spearman's r_s and Kendall's τ. A nonparametric test is preferred because as sample size decreases, the sampling variance increases (see fig. 1*A*), so the relation between effect size and sample size cannot legitimately be considered bivariate normal. These two nonparametric tests differ in the weighting of pairs of ranks. Spearman's r_s is preferred where the reliability of closely ranked values is uncertain (Sokal and Rohlf 1995, p. 600). Results from both tests are presented to confirm that statistical conclusions did not depend on the choice of test. Probability values for r_s and τ (r_{bias} and τ_{bias} in the table later) were two-tailed. Both were computed using Statview II (version 1.03, Abacus Concepts).

The correlation between effect size and sample size—one measure of the magnitude of selective reporting—is itself an "effect" (r_{bias}) amenable to standard meta-analytic methods. To compare the statistical significance of differences in r_{bias} between subsets of samples, I followed the procedure recommended by Rosenthal (1991): r_{bias} (based on Spearman's r_s) was converted to Fisher's $Z_r = 0.5 \log_e[(1 + r_{\text{bias}})/(1 - r_{\text{bias}})]$. Although normally applied to product-moment correlation coefficients, Fisher's Z_r may also legitimately be computed for Spearman's r_s where $n \geq 10$ and $\rho_s \leq 0.9$ (Zar 1984, p. 320). The significance of the difference between two estimates of r_{bias} was computed as $Z_{\text{diff}} = (Z_1 - Z_2)/\sqrt{[1.06/(N_1 - 3)] + [1.06/(N_2 - 3)]}$ (Zar [1984, p. 320] recommends 1.06 rather than 1 when using Z_r computed from r_s), which is distributed as t_s for ∞ df (Zar 1984, p. 313). Two-tailed probabilities were computed where no a priori direction of bias was expected, and one-tailed probabilities were used where specific biases were expected.

Weighted mean effect sizes at the level of samples were computed following Rosenthal (1991): $\overline{Z}_r = (\Sigma w_j Z_{rj})/\Sigma w_j$, where $w_j = N_j - 3$ (N_j = sample size for sample j) and Z_{rj} is Fisher's Z_r transformation of effect size (r) for sample j. Weighted mean effect sizes at the level of studies were computed similarly. For each study k, a value of Z_{rk} was computed in the same way as $\overline{Z}_r$ for all the samples in study k, and an average sample size computed as $N_k = (\Sigma N_j)/S_k$ (S_k = number of separate samples in study k). Across studies, the weighted mean effect was computed as $\overline{Z}_{rk} = (\Sigma w_k Z_{rk})/\Sigma w_k$, where $w_k = N_k - 3$. All values of Z_r were converted back to r to aid comparison ($r = [e^{2Zr} - 1]/[e^{2Zr} + 1]$; Rosenthal 1991, p. 71).

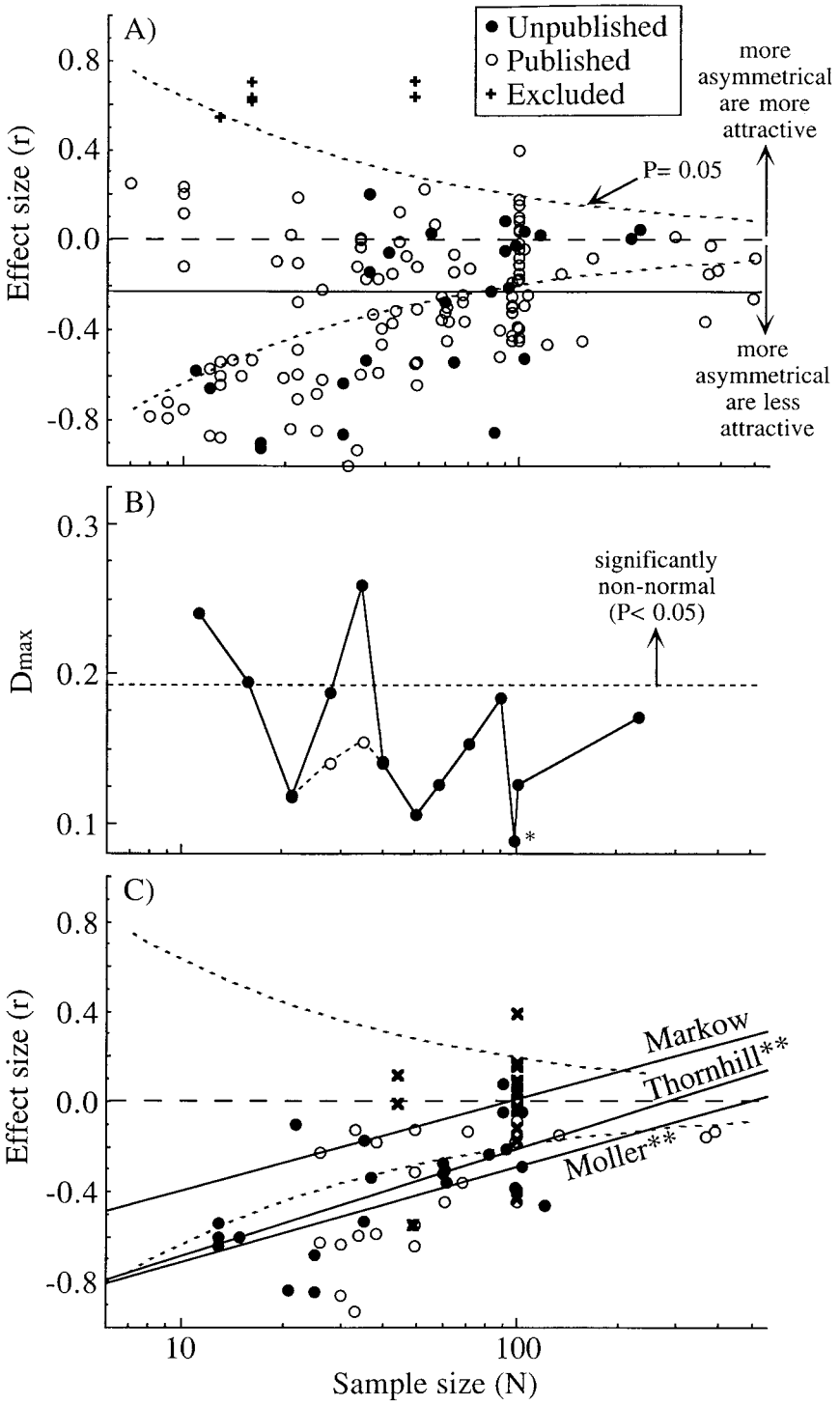

Figure 2: *A*, Effect size (r) as a function of sample size (N) for all samples included in table 1 of Møller and Thornhill (1998). *Solid circles*, unpublished samples; *open circles*, published samples included in the original analyses; *plus signs*, published samples excluded by Møller and Thornhill (1998). The solid horizontal line indicates the weighted mean for included samples ($N = 140$). *B*, Departure from normality (D_{max}; Kolmogrov-Smirnov test) as a function of sample size for the data presented in graph *A*, except for the excluded samples (effect size r was converted to Z_r before testing). The dashed line indicates the critical value ($\alpha = 0.05$) for $N = 20$ (table 33 of Rohlf and Sokal 1981). Open circles illustrate the value of D_{max} when the two samples, for which $r = -1.0$ had to be converted arbitrarily to $r = -0.99$ to compute Z_r, were excluded. The asterisk indicates the one group of 26 as opposed to 20 ± 2 samples. *C*, Effect size (r) as a function of sample size (N) for published and unpublished samples where Markow (*crosses*), Møller (*open circles*), or Thornhill (*solid circles*) were coauthors. Solid lines indicate least squares linear regression fit to the data and are for illustration only (see table 1, rows e–g, for statistics). Double asterisks indicate r_{bias} was significant statistically ($P < .01$). Note that the apparent slope for Markow is due entirely to the single observation at $N = 49$ and as a consequence does not differ significantly from 0 ($P > .6$). Short-dashed lines indicate where more extreme values of the statistics (r or D_{max}) become significant statistically ($P < .05$), and the long-dashed lines indicate an effect size of 0.

Table 1: Correlations between effect size (r) and sample size [$\log_{10}(N)$] for various subsets of cases from Møller and Thornhill (1998)

		Spearman rank correlation[a]		Kendall rank correlation[a]	
Samples included	$N_{samples}$	r_{bias}	P[b]	τ_{bias}	P[b]
a. All samples	146	.301	<.001***	.217	<.001***
b. All samples ("included" only)[c]	140	.393	<.001***	.280	<.001***
c. Unpublished samples	26	.663	<.001***	.495	<.001***
d. Published samples ("included" only)[c]	114	.345	<.001***	.248	<.001***
e. Markow	15	.110	.681	.095	.622
f. Møller[d]	24	.540	.010**	.412	.005**
g. Thornhill[d]	25	.650	.002**	.451	.002**
h. Møller and Thornhill[d]	49	.593	<.001***	.412	<.001***
i. Other authors[d] ("included" only)[c]	91	.312	.003**	.219	.002**
j. Møller (secondary sexual trait)[d]	10	.875	.009**	.719	.004**
k. Møller (ordinary trait)[d]	14	.073	.792	.047	.813
l. Thornhill (human face)[d]	10	.925	.006**	.809	.001**
m. Thornhill (human skeleton)[d]	9	−.020	.962	−.057	.830

Note: All significant results remain significant at $P < .05$ after a sequential Bonferroni correction is applied separately to each column of correlation coefficients.

[a] Corrected for ties.

[b] Two-tailed probability.

[c] Six samples excluded by Møller and Thornhill (1998) not included.

[d] Published and unpublished.

** $P < .01$.

*** $P \leq .001$.

Results

Excluded Samples

The studies excluded from the meta-analysis by Møller and Thornhill (1998) were the six highest positive effects out of all 146 studies (fig. 2*A*, *plus sign*). The likelihood that, owing to chance alone, the only six studies suffering from methodological or conceptual problems exhibited a positive effect (i.e., attractiveness increases with increasing asymmetry) is $P < .0001$ (contingency table analysis, χ^2 [corrected for continuity] = 17.06). That they should be the six most extreme positive values would be even less probable owing to chance.

Four of the excluded samples dealt with human facial asymmetry, but at least six other samples using artificially symmetrical human hemifaces were not excluded (Perrett et al., unpublished study; Mealey et al., unpublished study). One would have thought it more appropriate to exclude all 10 samples based on artificially symmetrical human hemifaces if their validity was in doubt or to explain more clearly why some hemiface methods are considered valid and others are not.

Tests for Selective Publication

Frequency Distribution of Effect Sizes. When effect size was viewed as a function of sample size (fig. 2*A*), several patterns emerged. First, as expected on purely statistical grounds, the range of effect sizes increased as sample size decreased. However, few clearly nonsignificant samples (those lying between the 95% significance thresholds of fig. 2*A*) were reported for sample sizes <20. As a consequence, frequency distributions of effect size at small sample size differed significantly from a normal distribution (fig. 2*B*).

Overall Dependence of Effect Size on Sample Size. When all 146 samples were included, r_{bias}—the correlation between effect size and sample size—was statistically significant ($P < .001$; table 1, row a). In other words, as sample size decreased, the "predicted" negative correlation between asymmetry and attractiveness became more pronounced (fig. 2*A*). The r_{bias} was even more pronounced when the six "excluded" samples were excluded (table 1, row b).

Comparisons of published and unpublished studies, where possible, are a recommended procedure in meta-analysis because unpublished studies should yield less biased estimates of effect size (Begg 1994). However, for the studies reported by Møller and Thornhill (1998), r_{bias} was also significant for unpublished samples ($P \leq .001$; table 1, row c). A weaker r_{bias} was observed among "included" published samples (table 1, row d), but r_{bias} was not quite significantly lower for published samples compared with unpublished ones ($P = .062$; table 2, row a).

Table 2: Tests for differences in the extent of r_{bias} between subsets of samples

Correlations compared	Rows in table 1	Z_{diff}	P	Test type
a. Unpublished vs. published[a]	c vs. d	1.868	.062	Two-tailed
b. Møller/Thornhill[b] vs. other authors[a,b]	h vs. i	1.938	.053	Two-tailed
c. Møller only:[b] sexual trait vs. ordinary trait	j vs. k	2.602	.005**	One-tailed
d. Thornhill only:[b] human face vs. human skeleton	l vs. m	2.912	.002**	One-tailed

Note: Both significant results remain significant at $P < .05$ after a sequential Bonferroni correction.

[a] Six samples excluded by Møller and Thornhill (1998) not included.

[b] Published and unpublished.

** $P < .01$.

Dependence of Effect Size on Sample Size for Individual Investigators. Only three investigators contributed enough samples ($N > 10$) to warrant individual tests of r_{bias}: Markow, Møller, and Thornhill. Although the least squares linear regression slopes were similar for all three (fig. 2*C*), the slope for samples from Markow resulted entirely from a single value and did not even approach statistical significance ($P > .6$; table 1, row e). In addition, only three of the 15 samples from studies by Markow were significant by themselves, so the results from Markow's studies (all on *Drosophila*) imply no overall significant association between asymmetry and attractiveness.

The r_{bias} was, however, highly significant for samples from Møller ($P \leq .01$; table 1, row f) and from Thornhill ($P < .002$; table 1, row g), as well as for both combined ($P < .001$; table 1, row h). The r_{bias} was also significant for samples from authors other than Møller and Thornhill ($P = .003$; table 1, row i), though it became marginally nonsignificant if the six excluded samples were included ($r_s = 0.195$, $P = .056$, $N = 97$). Compared with samples from all other authors, the samples from both coauthors of the original meta-analysis exhibited a higher r_{bias}, though this difference was not quite significant statistically ($P = .053$; table 2, row b). However, if the six excluded samples were included, r_{bias} became significantly greater among samples from Møller and Thornhill than from all other authors ($P = .004$).

To put values of r_{bias} into perspective, approximately 35% ($r^2_{bias} = 0.59^2$) of the variation in effect size among samples from Møller and Thornhill was due to sample size, whereas <10% ($r^2_{bias} = 0.31^2$) was due to sample size among samples contributed by all other authors.

Dependence of Effect Size on Sample Size for Particular Contrasts. Where sample sizes permit, tests for r_{bias} may also be conducted for particular contrasts within the studies of individual investigators. These may provide insights into how overall patterns of r_{bias} may have arisen. For example, among studies by Møller (fig. 3*A*), r_{bias} was significant for secondary sexual traits ($P = .009$; table 1, row j) but not for ordinary traits ($P = .8$; table 1, row k). Similarly, for studies on humans by Thornhill (fig. 3*B*), r_{bias} was significant for faces ($P = .006$; table 1, row l) but not for skeletal traits ($P > .95$; table 1, row m). In both cases, r_{bias} was significantly higher for the trait predicted by the authors to show an effect of asymmetry on attractiveness ($P \leq .005$; table 2, rows c, d).

Correction for Multiple Tests. Note that all of the statistically significant correlations in table 1 remain significant after a sequential Bonferroni correction for multiple tests (Rice 1989) is applied separately to each column of correlation coefficients. Because the two correlation coefficients (r_{bias} and τ_{bias}) yield virtually identical results, and the results for both were included only to illustrate this point, the Bonferroni correction is most appropriately applied to the results for each coefficient separately (i.e., correct for 13 tests) as opposed to all combined (26 tests). In addition, both of the statistically significant correlations in table 2 remain significant after a sequential Bonferroni correction.

Impact of Selective Publication on Estimates of Effect Size and Effect Size Heterogeneity

Overall Effect Size. I was unable to reproduce some of the statistical descriptors reported by Møller and Thornhill (1998). For example, I could not reproduce the weighted mean effect size for included samples ($r = -0.42$; row 2 in their table 2), even though I did reproduce the weighted mean effect size for the six excluded samples ($r = 0.65$; row 1 in their table 2). For the included samples, I obtained a weighted mean effect size of $r = -0.229$ ($N = 140$; and for all 146 samples $r = -0.215$). To compute this value for all 140 samples, however, two effect sizes of $r = -1.0$ had to be adjusted (I arbitrarily chose $r = -0.99$), since as $r \to \pm 1.0$, $Z_r \to$

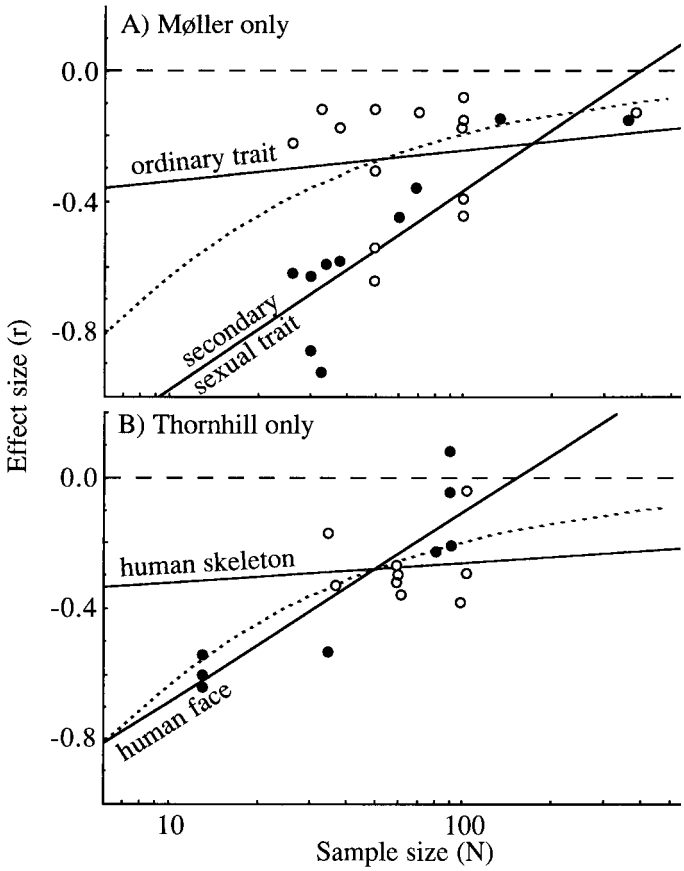

Figure 3: Effect size (r) as a function of sample size (N) for published and unpublished samples for individual investigators. Solid lines indicate least squares linear regression fit to the data (for both figures, the line with the lower slope applies to the open circles). See figure 2 legend for an explanation of the dashed lines. *A*, Samples in which Møller was a coauthor (see table 1, rows j, k, for statistics); *B*, samples in which Thornhill was a coauthor (see table 1, rows l, m, for statistics).

$\pm\infty$ (Zar 1984, p. 310). If Møller and Thornhill substituted a more extreme value (e.g., -0.99999), their estimate of weighted mean effect size would have been inflated. Furthermore, for these data, the weighted mean must be less extreme (less negative) than the unweighted mean ($r = -0.299$, $N = 140$) because the more extreme r values at small sample sizes would have contributed less. Finally, the distribution of effect sizes in figure 2*A* is not visibly consistent with a weighted mean of $r = -0.42$.

In addition, I was also unable to reproduce some weighted means reported by Møller and Thornhill (1998) for the study level of analysis. I obtained a weighted mean effect of $r = -0.223$ (all 64 unique studies listed in their table 1), or $r = -0.239$ (for the 61 studies remaining after the six excluded samples were removed). These weighted mean effect sizes are considerably lower than those reported in their table 2 (-0.36 and -0.42 in rows 4 and 3, respectively).

Møller and Thornhill appear to have overestimated the overall mean effect size at the level of both samples and studies by nearly twofold.

Effect Size Differences for Particular Contrasts. Significantly, if the weighted mean effect size for all included studies is $r = -0.239$ ($N = 61$), then the weighted mean effect sizes tabulated for various contrasts of interest (rows 3–12 in table 2 of Møller and Thornhill 1998) must also be too high. For example, if one weighted mean from a pair of effects in a contrast is above the grand mean, the other must be below it, yet nearly all the values tabled in rows 3–12 of their table 2 are more extreme than the weighted mean I computed of $r = -0.239$.

Computational discrepancies aside, a graphical approach suggests that quantitative contrasts of effect size statistics may be potentially misleading where selective reporting appears widespread (fig. 4). For three contrasts, cases of small sample size ($N \leq 30$) and large effect size were more numerous where correlations between FA and

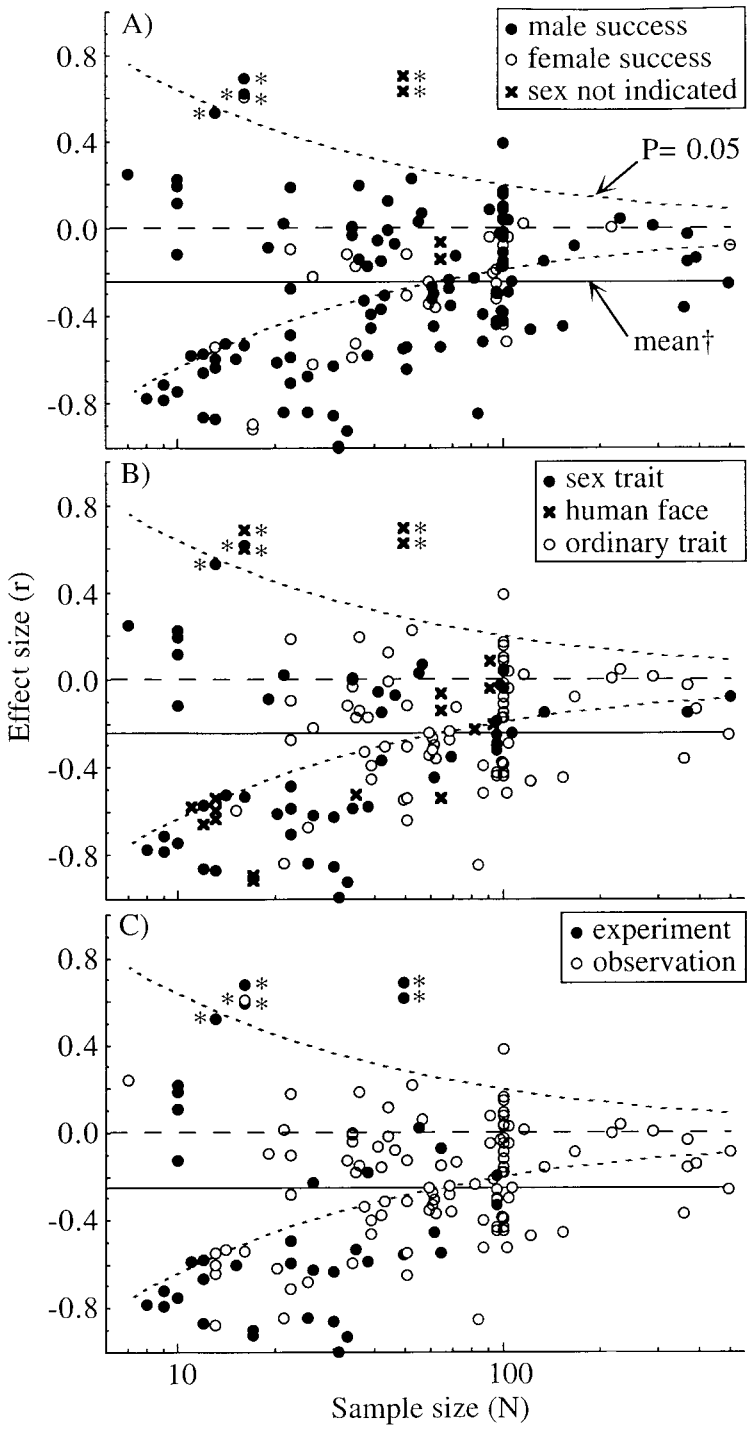

Figure 4: Effect size (r) as a function of sample size (N) for three contrasts of the effect of asymmetry on sexual selection. All effect sizes are from table 1 of Møller and Thornhill (1998). Asterisks indicate samples excluded from the analyses by Møller and Thornhill; the dagger indicates the weighted mean effect size for the 140 included samples. *A*, Effect of asymmetry on the mating success of males compared with females; *B*, effect of asymmetry on mating success for three types of traits; *C*, effect of asymmetry on mating success as determined by experimental or observational studies.

attractiveness were expected to be more pronounced. Reports of male success were proportionally more common than those of female success among cases based on small (83%) versus large (76%) sample size (fig. 4*A*), reports of sex traits (including human face) were proportionally more common than those of ordinary traits among cases of small (80%) versus large (35%) sample size (fig. 4*B*), and reports from experimental studies were proportionally more common than those from observational studies among cases of small (55%) versus large (13%) sample size (fig. 4*C*). These proportions did not differ significantly for studies of male compared to female success ($\chi^2 = 0.45$, $P = .50$; χ^2 test with correction for continuity). They were, however, highly significantly different for sex versus ordinary traits and experimental versus observational results ($P < .001$). As a consequence, the effects of selective reporting seriously confound the computations of mean effect size for these latter two contrasts.

Impact of Particular Investigators on Estimates of Mean Effect Size. In view of the significant r_{bias} among samples contributed by Møller and Thornhill (figs. 2*C*, 3*A*, *B*; table 1, row h; table 2, row b), one might wonder to what extent their own contributions influenced the estimate of overall effect size. For all included samples (published and unpublished), the weighted mean effect at the level of studies by authors other than Møller and Thornhill was $r = -0.217$ ($N = 44$). Therefore, although the contributions by Møller and Thornhill amplify the mean effect size at the level of studies by about 10% (-0.239 vs. -0.217), the overall effect did not depend on their contributions.

Discussion

Meta-analysis offers several advantages over narrative summaries of the literature (Arnqvist and Wooster 1995). In particular, it requires that results be presented as a standardized statistic—effect size—so they are more readily comparable. In the present study, "effect size" refers simply to the strength of the correlation (r) between asymmetry and attractiveness: a negative effect size means that attractiveness decreases as asymmetry increases. Numerical values for effect size may be computed from a variety of statistics reported in the original studies using standard meta-analytic procedures (Rosenthal 1991).

In addition to standardizing results, meta-analysis provides tools for computing an average effect size across multiple studies as well as formal statistical methods for detecting heterogeneity among effect sizes and asking whether particular contrasts between effects of interest may contribute to that heterogeneity. These are valuable applications. However, because selective reporting may introduce unwanted biases, exploratory analyses of reporting patterns, particularly as they relate to sample size, should be conducted routinely before computing quantitative estimates of effect size and effect size heterogeneity (Begg 1994).

The funnel graph approach of Light and Pillemer (1984), and its quantitative extensions, offer powerful tools for detecting selective reporting and therefore help to encourage caution when computing average effect sizes or interpreting patterns of effect size variation. When applied to the extensive collection of studies of FA and sexual selection (Møller and Thornhill 1998), the funnel graph approach offers some sobering revelations. First, selective reporting appears to be widespread and to have inflated estimates of overall effect size. Second, selective reporting may have confounded tests of effect size differences for contrasts of both biological and methodological interest.

Evidence of Selective Reporting

As noted earlier, selective reporting in meta-analyses may be signaled in two ways (Light and Pillemer 1984), and both signals (fig. 1*B*, *C*) were apparent among the studies of FA and sexual selection (Møller and Thornhill 1998). First, figure 2*A* resembles the funnel-shaped pattern expected owing to selective reporting (fig. 1*B*), and the significant departure from normality of effect sizes for sample sizes ≤ 20 (fig. 2*B*) supports such a conclusion. In addition, the closeness with which the lower 95% significance threshold delimits the upper edge of the cluster of negative effects at small sample size ($N \leq 20$, lower left portion of fig. 2*A*) suggests that statistical significance had a strong impact on likelihood of publication. Second, when all effect sizes were examined together, whether published or unpublished, the dependence of effect size on sample size (r_{bias}) was significant statistically (table 1, rows a, b).

A significant r_{bias} among unpublished cases (fig. 2*A*; table 1, row c) was particularly surprising. In general, because unpublished studies should be more representative of all studies conducted (Light and Pillemer 1984; Begg and Berlin 1988), r_{bias} should be less pronounced. However, since these unpublished results may have been from manuscripts volunteered by other authors who had prepared them for publication, they may not have been a representative sample of unpublished studies. Nonetheless, the significant r_{bias} suggests that authors were more likely to provide Møller and Thornhill with results that were consistent with expectations (i.e., a statistically significant negative correlation between asymmetry and attractiveness).

Finally, r_{bias} also varied in a surprising way in relation to a priori expectations of association between asymmetry and attractiveness among studies by individual authors. Such analyses are informative because they are not con-

founded by among-author variation. Unfortunately, this could only be examined rigorously for reports from Møller and Thornhill themselves, where a sufficiently large and roughly equivalent number of samples permitted a more detailed analysis.

The most forceful prediction of several advanced by Møller and Thornhill (1998 and references therein) is that asymmetry of secondary sexual traits should have a greater negative effect on attractiveness than asymmetry in ordinary traits. For this test, they found strong statistical support, particularly when the results for human faces were included (rows 7, 8 in their table 2). However, when their own studies were examined in more detail (fig. 3*A*, *B*), r_{bias} was highly significant for secondary sexual traits ($P < .001$; table 1, rows j, l) but not ordinary traits ($P > .7$; table 1, rows k, m). This same pattern was also evident among all samples (fig. 4*B*).

Because investigator effects are inevitable, they need not imply differences in the quality or validity of studies from different labs: different investigators legitimately use different methods in different systems. Therefore, for example, the greater effect sizes reported in studies by Møller and Thornhill (fig. 2*C*) may simply reveal that they are better than others at choosing systems or methods of study that reveal the "true" effects of asymmetry on attractiveness. The differences in r_{bias} between signaling and nonsignaling traits among their own studies (fig. 3*A*, *B*), however, remain puzzling.

Impact on Estimates of Overall Effect Size

Evidence of selective reporting is neither surprising nor original (Cooper and Hedges 1994), and I do not intend to imply that it is somehow more prevalent among studies of FA and sexual selection than elsewhere. As noted earlier, selective reporting, and the bias it introduces, seems an unavoidable consequence of the research enterprise (Begg and Berlin 1988; Begg 1994). The question is, Does selective reporting invalidate or weaken conclusions about either statistical or biological significance?

The fail-safe criterion (Rosenthal 1979) provides a rough idea of how great concerns should be about the statistical impact of selective publication. The revised overall effect size computed from Møller and Thornhill (1998) for included studies ($r = -0.239$, $N = 61$) is unlikely to be due solely to selective publication (the file drawer problem; Rosenthal 1979): 8,000 or more studies of zero effect (the fail-safe number of studies; Cooper 1979) would have to be published to eliminate the statistical significance of this result. For the 41 studies based on larger sample sizes ($N > 30$), the fail-safe number is smaller (3,742), though still sizable; and for the 44 studies contributed by authors other than Møller and Thornhill, it was also comparable (3,383). Therefore, in spite of direct evidence for selective reporting (figs. 2, 3), roughly 75–100 unpublished studies of zero effect would have to exist per published study to reduce the overall effect to nonsignificance. However, as noted above ("Interpreting Results in the Presence of Selective Reporting"), if effects of opposite (positive, in this case) sign were common, far fewer would be needed to reduce the average effect size to nonsignificance.

The biological significance of the overall association between FA and attractiveness, however, seems less clear. The revised estimates of overall effect size reported here, both at the level of samples ($r = -0.229$, $N = 140$) and the level of studies ($r = -0.239$, $N = 61$), are substantially lower than those originally reported by Møller and Thornhill (1998). Both revised estimates are much closer to the effect size of $r = -0.26$ computed independently by Leung and Forbes (1996) for the relation between FA and measures of fitness. Regardless of the statistical significance of these estimates, if they are approximately correct, they imply that across all studies conducted to date, <6% of the variation in rank scores of attractiveness ($r^2 = -0.239^2$) can be attributed to variation in FA. Furthermore, because of the publication biases noted earlier, this is surely an overestimate.

Contrary to the conclusions of Møller and Thornhill (1998), and statistical significance notwithstanding, if we accept these revised estimates of average effect size at face value, FA appears to account for surprisingly little (<6%) of the variation in attractiveness. The biological significance and generality of this association therefore seem to have been greatly overstated.

Impact on Estimates of Effect Size Heterogeneity

Whether selective reporting affects conclusions about differences in effect size among groups depends on whether it varies in any systematic way for contrasts of biological or methodological interest. For example, if effects for one subgroup (e.g., males) are often based on smaller sample sizes than a comparison subgroup (e.g., females), then the biases due to selective publication artificially inflate the effect size difference between these subgroups.

Inspection of funnel graphs suggests that conclusions drawn from two contrasts examined by Møller and Thornhill (1998) may have been overstated. First, effect sizes for secondary sexual traits (including the human face) formed a majority of cases based on sample sizes of 30 or fewer (32 of 40), whereas effect sizes for ordinary traits formed a majority of cases based on sample sizes greater than 30 (65 of 100; fig. 4*B*). Furthermore, for cases based on sample sizes greater than 40, figure 4*B* reveals no apparent difference in effect size for secondary sexual traits compared with ordinary ones. Therefore, the conclusion that

the effect of asymmetry on attractiveness is more pronounced for secondary sexual traits than ordinary ones depends heavily on results based on small sample sizes, and it does not appear supported by studies based on larger sample sizes (fig. 4*B*).

Second, effect sizes from experimental studies formed a majority of cases based on sample sizes of ≤30 (22 of 40), whereas those for observational studies formed a majority of cases based on sample sizes >30 (87 of 100; fig. 4*C*). Although the statistical consequences of this pattern are the same as noted earlier for secondary sexual versus ordinary traits, the interpretation is less clear. Controlled experimental studies may not require as large a sample size as observational studies to detect significant differences. Investigators may intentionally plan experimental studies with smaller sample sizes than observational ones for just this reason, so the dependence of effect size on data type may represent rational planning more than selective publication. However, the closeness with which the 95% significance threshold defines the upper limit to the cluster of samples in the lower-left corner of figure 4*C* suggest that, even among experimental studies, statistical significance of an outcome had a strong effect on likelihood of publication. Therefore, the conclusion that experimental studies yield greater effect sizes than observational ones (Møller and Thornhill 1998) may also have been overstated.

Conclusions

Although some may question the validity of meta-analyses conducted on studies from a highly heterogeneous set of taxa and traits, meta-analysis nonetheless offers numerous advantages over narrative summaries of the literature, including the ability to evaluate the impact of selective reporting. Because they are informative and easy to conduct, both graphical and nonparametric tests for selective reporting would seem profitable to incorporate as routine components of meta-analyses. As Light and Pillemer (1984) note so pointedly: "Taking an average across a series of outcomes is rarely a difficult conceptual issue. The difficult question is how to treat differences among the findings that invariably turn up—the variability of different outcomes about the average. Anyone summarizing 15, or 50, or 100 results with one statistic must face a fact: the cost of using a simple summary index is a loss of information. How one views that fact has important consequences" (p. 51).

In contrast to statistical summaries, graphical methods (e.g., fig. 4) provide far more useful insights into patterns of publication and their possible impact on conclusions about average effect sizes and about effect size heterogeneity.

An application of these graphical techniques revealed that selective reporting (as r_{bias}) appears to be widespread in studies of FA and sexual selection. It was noticeable among both published and unpublished studies. It was more pronounced among studies conducted by the authors of the original meta-analysis (Møller and Thornhill) than among other authors, and among studies by these two authors, it was more pronounced where correlations with asymmetry were predicted by them to be stronger, such as for secondary sexual traits compared to ordinary traits. Finally, among all studies, the prevalence of large effects among studies of small sample size for secondary sexual traits suggests that conclusions about differences in the correlation between asymmetry and attractiveness for secondary sexual traits compared with ordinary traits are likely overstated.

For phenomena such as fluctuating asymmetry, where the acknowledged biological signal is so exceedingly small, we must guard against being deceived by statistical oversimplification. Studies based on sample sizes of <20 or 30 seem the most prone to bias, so perhaps reviewers and editors should ask for significance levels of .01 or .001 for studies based on small sample sizes to reduce the impact of selective reporting.

Acknowledgments

I thank D. Arsenault, L. Hammond, J. Kingsolver, C. Klingenberg, D. Repasky, and C. Strobeck for their helpful comments on the manuscript and L. Rimmer of the Bamfield Marine Station Library for help obtaining locally unavailable references. I am particularly grateful to three anonymous reviewers more familiar with meta-analysis than I for their detailed and constructive suggestions regarding organization, terminology, and emphasis and for direction to more recent references on publication bias. This research was supported by Natural Sciences and Engineering Research Council of Canada operating grant A7245.

Literature Cited

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

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

Begg, C. B. 1994. Publication bias. Pages 399–409 *in* H. Cooper and L. V. Hedges, eds. The handbook of research synthesis. Russel Sage Foundation, New York.

Begg, C. B., and J. A. Berlin. 1988. Publication bias: a

problem in interpreting medical data. Journal of the Royal Statistical Society A151:419–463.

Cooper, H. 1979. Statistically combining independent studies: a meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology 37:131–146.

Cooper, H., and L. V. Hedges, eds. 1994. The handbook of research synthesis. Russel Sage Foundation, New York.

Houle, D. 1997. Comment on "A meta-analysis of the heritability of developmental stability" by Møller and Thornhill. Journal of Evolutionary Biology 10:17–20.

Iyengar, S., and J. B. Greenhouse. 1988. Selection models and the file drawer problem. Statistical Science 3: 109–135.

Leamy, L. 1997. Is developmental stability heritable? Journal of Evolutionary Biology 10:21–29.

Leung, B., and M. R. Forbes. 1996. Fluctuating asymmetry in relation to stress and fitness: effects of trait type as revealed by meta-analysis. Ecoscience 3:400–413.

Light, R. J., and D. B. Pillemer. 1984. Summing up: the science of reviewing research. Harvard University Press, Cambridge, Mass.

Møller, A. P. 1997. Developmental stability and fitness: a review. American Naturalist 149:916–932.

Møller, A. P., and J. P. Swaddle. 1997. Developmental stability and evolution. Oxford University Press, Oxford.

Møller, A. P., and R. Thornhill. 1997. A meta-analysis of the heritability of developmental stability. Journal of Evolutionary Biology 10:1–16.

———. 1998. Bilateral symmetry and sexual selection: a meta-analysis. American Naturalist 151:174–192.

Palmer, A. R. 1996. Waltzing with asymmetry. BioScience 46:518–532.

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223–225.

Rohlf, F. J., and R. R. Sokal. 1981. Statistical tables. W. H. Freeman, San Francisco.

Rosenthal, R. 1979. The "file drawer problem" and tolerance for null results. Psychological Bulletin 86: 638–641.

———. 1991. Meta-analytic procedures for social research. Sage, Beverly Hills, Calif.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. 3d ed. W. H. Freeman, New York.

Vevea, J. L., and L. V. Hedges. 1995. A general linear model for estimating effect size in the presence of publication bias. Psychometrika 60:419–435.

Zar, J. H. 1984. Biostatistical analysis. 2d ed. Prentice Hall, Upper Saddle River, N.J.

Associate Editor: Joel G. Kingsolver