Annu. Rev. Ecol. Syst. 2000. 31:441–80

QUASIREPLICATION AND THE CONTRACT OF ERROR: Lessons from Sex Ratios, Heritabilities and Fluctuating Asymmetry

A. Richard Palmer
Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9 Canada, and Bamfield Marine Station, Bamfield, British Columbia V0R 1B0 Canada; e-mail: rich.palmer@ualberta.ca

Key Words publication bias, selective reporting, funnel graph, research synthesis, replication, meta-analysis

■ **Abstract** Selective reporting—e.g., the preferential publication of results that are statistically significant, or consistent with theory or expectation—presents a challenge to meta-analysis and seriously undermines the quest for generalizations. Funnel graphs (scatterplots of effect size vs. sample size) help reveal the extent of selective reporting. They also allow the strength of biological effects to be judged easily, and they reaffirm the value of graphical presentations of data over statistical summaries.

Funnel graphs of published results, including: (*a*) sex-ratio variation in birds, (*b*) field estimates of heritabilities, and (*c*) relations between fluctuating asymmetry and individual attractiveness or fitness, suggest selective reporting is widespread and raise doubts about the true magnitude of these phenomena. Quasireplication—the "replication" of previous studies using different species or systems—has almost completely supplanted replicative research in ecology and evolution. Without incentives for formal replicative studies, which could come from changes to editorial policies, graduate training programs, and research funding priorities, the contract of error will continue to thwart attempts at robust generalizations.

> "For as knowledges are now delivered, there is a kind of contract of error between the deliverer and the receiver: for he that delivereth knowledge desireth to deliver it in such a form as may be best believed, and not as may be best examined; and he that receiveth knowledge desireth rather present satisfaction than expectant inquiry; and so rather not to doubt than not to err: glory making the author not to lay open his weakness, and sloth making the disciple not to know his strength."
>
> *The Advancement of Learning*, Francis Bacon, 1605 (8:170–171)

INTRODUCTION

Little has changed since Bacon penned these perceptive words nearly four hundred years ago. He clearly recognized a fundamental weakness of human nature: We prefer entertainment to challenge. This weakness—when coupled with a deep faith in modern statistics, a general unwillingness to acknowledge that some results will appear significant due to chance, and journal editorial policies that explicitly discourage replication—potentially undermines our quest for a robust understanding of ecological and evolutionary phenomena. Together, these conspire to perpetuate a collective contract of error, where popular beliefs are sanctified by the selective publication of results that are either statistically significant or consistent with theory or expectation, and where a peer-review process discourages the very contributions that are needed most: formal replicate studies.

How do we come to accept that a generalization has been demonstrated scientifically? For the most part, we judge a generalization's validity based on our reading of the scientific literature and our sense of the internal consistency of this literature and its consistency with our own personal observations. But if statistical significance of a result—or its concordance with pre-conceptions—influences the likelihood of publication, or if oft-repeated claims in review papers reinforce belief in a paradigm even where the evidence remains ambiguous (33), then how many emerging generalities reflect biological reality as opposed to collective wishful thinking embroidered with statistical support?

This problem is exacerbated when advocates buttress their strongly held vision of a particular phenomenon with seemingly compelling theoretical and empirical support (42, 73), thereby encouraging others to follow along with their own independent confirmations (2, 86, 96). Eventually, if the claims are exaggerated, either (*a*) the primary protagonists retire or pass on, and without their continued proselytizing others lose interest (as eventually happened to one branch of quantitative genetic methods, (70), or (*b*) the claims trigger a backlash among skeptics, and the weight of evidence declines or swings in the opposite direction (e.g., see Figure 13 below, and Refs. 2, 84, 86, 96, as possible examples). Such oscillations can yield sufficient distrust that even intriguing biological phenomena may be ignored because the waters have become so muddied by contradictory claims.

In fields where repeated tests of specific hypotheses are more commonplace (physics, molecular biology, medicine), an average result seems eventually to emerge from the variation among studies. This happens, in part, because the success of subsequent work depends so much on replication of prior methods and results. In ecology and evolution, where particular outcomes are typically repeated with different species or systems rather than truly repeated, we build up an impression of repeatable patterns by averaging over many heterogeneous studies either qualitatively or else quantitatively via meta-analysis. But, as will become apparent, these tests are as likely to reinforce an illusion as they are to validate a biological pattern.

Quasireplication and Selective Reporting

The practice of repeating studies with different species or systems—but not with the same species or system—seems so entrenched in many biological disciplines, including ecology and evolutionary biology, that I think it deserves a name: *quasireplication*. Quasireplication refers to what others have variously called, in different contexts, "imitative or acquisitive study" (18), "normal science" (60), "encyclopedism" (85), "advocacy science" (120), or "corroborative research" (101). Unfortunately, quasireplication is not true replication. Unless it is combined with true replication, it seems more likely to mislead than to reassure for two reasons.

First, quasireplications seem more vulnerable to selective reporting (the tendency to publish only a subset of studies that were undertaken) and therefore likely to lead to publication bias (deviations in average effect sizes caused by selective reporting) (12, 13). Authors who have explicitly set out to replicate a previous study as closely as possible will likely have more at stake in the outcome, since any outcome—whether confirmatory or not—would be of value. However, authors simply asking whether a popular pattern exists in their favorite organism or system may be less inclined to report negative or contradictory results unless they have unusually great confidence in the power of their test. Quasireplication may seriously compromise the validity of generalizations because repeated reports of popular or trendy results in a variety of different species or systems reinforce belief in their generality, even though much of the support may derive purely from selective reporting.

Second, quasireplication does not provide nearly the same strength of test as does true replication, and therefore it is less effective at resolving differences of opinion about the validity of hypotheses or results. For example, if quasireplicated studies present discrepant results, opponents of a particular hypothesis will highlight them as contradictory evidence, but proponents of the same hypothesis can simply dismiss them as not a true replication of the original claim. The result is bickering, hand waving, and meta-analyses conducted or interpreted in different ways to support one or another cherished belief (74, 80, 105). Such debate is almost entirely deflated by a few well-conducted, fully replicated tests of definitive or classical studies.

Quasireplication nonetheless has an important role. It offers the quickest route to true biological generalizations. However, it should not be used as a substitute for true replication, because of its vulnerability to selective reporting. Clearly, what is needed is a proper balance between the two.

Prior Controversies

Controversies over the validity of prevailing dogma are not new. In the early 1980s, a veritable donnybrook erupted between community ecologists who believed, following Platt (85), that explicit hypothesis testing was the only proper research protocol (95, 101, 102), and others who argued against such a rigid approach (87) or

advocated a compromise (92, 93). Elner & Vadas (33) offered a particularly illuminating retrospective of the debate over mechanisms driving the lobster/urchin/kelp system in the western Atlantic, in which "over a period of approximately 20 years, [published] explanations for the phenomenon invoked four separate scenarios, which changed mainly as a consequence of extraneous events rather than experimental testing." Changes in viewpoint were driven as much by sociological as by biological factors. Although much of the debate revolved around how hypotheses are best tested, one recurring cry was for more formal replication of prior research rather than the accumulation of more quasireplicated studies (95, 101).

We seem to have made little progress in the intervening 20 years, but signs of change are increasingly evident. Concerns have recently been raised about selective reporting for several phenomena, including the adaptive significance of enzyme polymorphism (16), correlations with allozyme heterozygosity (52), adaptive sex-ratio variation in mammals (36), and relations between fluctuating asymmetry and sexual selection (80). Concerns have also been raised about selective reporting and publication bias in biological research in general (26). Apparently we are beginning to acknowledge the immensity of the problem.

Unfortunately, while meta-analytic techniques may help reveal biases due to selective reporting, they cannot, at present, reliably correct for them (12, 14, 19, 57). Furthermore, simple graphical techniques like the funnel graph (64) may prove more useful than summary statistics for judging the validity of emerging generalizations.

RESEARCH SYNTHESIS AND PATTERNS OF REPORTING

Few would dispute the value of quantitative approaches to reviews of the literature. But do such quantitative syntheses further exacerbate the contract of error by lending statistical support to collective preconceptions? Disentangling true biological effects from the biases introduced by selective reporting remains one of the most serious unsolved problems of meta-analysis (12, 13, 19, 50, 57).

Meta-Analysis: A Brief Overview

Formal methods of quantitative research synthesis form the domain of meta-analysis (25, 55, 77, 91). In meta-analysis, various statistical results from multiple studies are first converted to a standard statistic—effect size—to allow quantitative summarization. Several effect size statistics are available (25, 77, 91), but the correlation coefficient is perhaps the most popular in ecological and evolutionary studies (7).

Correlation coefficients simply describe the consistency of an association between two variables and have four advantages: (*a*) They are a familiar statistic; (*b*) they range between zero and ± 1.0; (*c*) when squared they yield the coefficient of determination (98), which describes the percent of variation in Y explained

by variation in X; and (*d*) statistical significance thresholds provide a useful reference against which published statistics may be compared to test for selective reporting (80). However, correlation coefficients reveal nothing about the strength of an association (i.e. the amount of change in Y for a given change in X), so the familiar—but often ignored—caution about not mistaking statistical significance for biological significance must be repeated.

Meta-analysis has a long history in psychological and educational research (45, 57, 68) and is now widely used to assess the strength and validity of medical research findings (57, 75, 77) [see Becker et al (11) for a particularly useful compact review of books on meta-analytic methods]. In these fields, concerns about publication biases have also been widespread (29, 61) because the validity of a quantitative outcome may have a profound and potentially expensive impact on public policy.

Meta-analysis has also received increasing attention in ecology and evolution following its original application to the effects of age on fecundity in birds (56), a prominent application to the great debate of the mid-1980s over the impact of competition in the field (45), and a convenient, compact review that introduced it to a wide audience (7). More recently, even Bayesian statistical approaches are being applied as an alternative to conventional meta-analytic methods (108). As the popularity of meta-analysis has increased, though, more and more authors have expressed concern about the impact of publication bias on statistical summaries (1, 2, 7, 80), a point that was not always emphasized in earlier meta-analyses (e.g., see 45, 47). Because of the profound bias selective reporting may introduce, great care must be taken to avoid simply re-enforcing the bias by relying on simplified summary statistics within a meta-analysis.

Although not without its shortcomings (27, 30, 39), meta-analysis offers a significant advance over narrative research synthesis. Just as cladistic methods have revolutionized phylogenetics—by forcing all steps in an analysis to be made transparent (including data to be analyzed, characters and character state definitions, weighting protocols, and analytical procedures)—meta-analysis has the potential to revolutionize research synthesis in ecology and evolution. Unfortunately, just as cladistic analyses can be slanted in a preferred direction, so can meta-analysis (31). The exchanges between Givens et al (43) and commentators (14) on the second-hand smoke debate, and between Palmer (80) and Thornhill et al (105) regarding fluctuating asymmetry and sexual selection, are particularly illuminating in this regard.

Potential Causes of Selective Reporting

Many factors influence probability of publication. Some are of little consequence to research synthesis (e.g., loss of funding, loss of motivation or interest unrelated to early results, distraction by other activities) because they are unrelated to a research outcome. However, other causes of underreporting—statistical nonsignificance, inconsistency with expectation, inconsistency with theory—may seriously bias

meta-analytic summary statistics. What data exist suggest this problem may be profound.

Statistical Significance of Results Perhaps the most familiar and widespread cause of selective reporting is the statistical significance of results (12, 13, 50). It may influence both an author's willingness or desire to press forward with publication and the willingness of editors and referees to accept a result.

The impact of statistical significance is most easily assessed when the outcomes of both published and unpublished studies may be evaluated retrospectively (12). In retrospective surveys of studies known to have been conducted, those yielding statistically significant results ($P < 0.05$) were more likely to be published, and published sooner (e.g., 24, 100). In addition, among nonsignificant studies, those yielding clearly nonsignificant results ($P > 0.1$) were more likely to be published than those yielding ambiguous results ($0.05 < P < 0.1$) (100). Clearly, not only the statistical significance, but also the statistical clarity of a study's outcome can influence its likelihood of publication.

Consistency with a Preferred Hypothesis Weighted mean effect sizes from a meta-analysis might be trusted if the effects of preconceptions, both positive and negative, averaged out. But what if a preconception is widely shared? Even if it had no biological validity it could still bias the weight of published evidence in its favor, and summary statistics from meta-analyses would serve only to reinforce this bias.

An example from medical research suggests just such a bias. When studies of the effects of acupuncture were compared to other randomized or controlled trials, certain countries reported disproportionately more positive findings than others (112). Those countries reporting more positive findings also happened to be countries in which acupuncture was considered an acceptable treatment. In addition, "no trial published in China or Russia/USSR found [an acupuncture] treatment to be ineffective" (112). Alternatively, in view of the well-known and sizeable placebo effect (49), perhaps the higher incidence of positive findings reflected cultural differences in the belief in acupuncture's effectiveness rather than selective reporting. Patients in China and Russia may, in fact, have shown demonstrably better responses to treatment.

Consistency with Theory Certain results might be under-reported because they make no sense theoretically, even though sampling error dictates that theoretically nonsensical values should arise occasionally due to chance.

Heritability estimates provide a test for such a bias. In the absence of sampling error, theory predicts that heritabilities should range from zero (no resemblance between parents and offspring) to 1.0 (offspring exactly resemble the mean phenotype of their parents) (35). For well-behaved polygenic traits, heritabilities in excess of 1.0 would imply that offspring consistently deviated more from the population mean than the average phenotype of their parents, and heritabilities less

than 0.0 would imply that the more parents deviated from the population mean in one direction the more their offspring deviated in the other!

Few evolutionary biologists would seem likely to place much belief in the biological significance of such extreme heritability estimates ($h^2 < 0.0$, $h^2 > 1.0$). However, sampling error dictates that some should arise simply due to chance, particularly when based on small sample sizes (66). The underreporting of negative heritabilities (see Heritability: Impact of Theoretical Preconceptions, below) reveals that consistency with theory clearly influences likelihood of publication.

Even where sampling error may be negligible, inconsistency with a strongly held theory may discourage acceptance of a result. For example (42a), prior to 1956, physicists believed that parity (i.e., mirror image counterparts of all physical phenomena including positive/negative charge, matter/antimatter, right/left spin) was always conserved. This belief was so strong that when three physicists observed a parity violation in the decay of a radium isotope in 1928 and again in 1930, the result was ignored even though it was seen "in all readings in every run, with few exceptions"(42a:218). The result simply did not coincide with any accepted theory. Not until 1956 was theory revised to admit the possibility of parity violations in weak interactions, for which the authors later received the Nobel Prize in physics. Shortly afterwards the first supposedly definitive violation of parity was seen in the beta decay of cobalt 60 (an excess of electrons is emitted from one end of the spinning nucleus) and then quickly confirmed for many other weak interactions. Inconsistency with theory had therefore discouraged the acceptance of repeatable observations of parity violation for over 25 years.

The "Fail-Safe" Number and Its Limitations

The fail-safe number (23, 91) is often invoked to reassure meta-analysts that selective reporting would have to have been severe to account for the overall statistical significance of a particular effect. It estimates the number of studies of zero effect that would have to be published to reduce a weighted-mean effect size to nonsignificance. A fail-safe number of 1000 therefore means that 1000 studies of zero effect would have to have gone unpublished—left in a file drawer (90)—for a particular average effect to have reached statistical significance due to selective reporting.

Although easy to compute and statistically well-defined, the fail-safe number can yield a deceptive impression of how robust a particular meta-analytic result is. This deception arises because the fail-safe computation assumes that all unpublished studies are of zero effect. Clearly, though, some unpublished studies must have yielded results in the opposite direction (e.g., compare Figure 7 to Figure 14 below). So while 1000 studies of zero effect might reduce a meta-analytic mean to nonsignificance, only 100 or fewer studies of zero and opposite effect could reduce a meta-analytic mean to nonsignificance. Therefore, "if the literature is one in which a large number of unreported studies with opposing results may exist, then the usual fail-safe number may add unwarranted confidence to the interpretation of the reported (but potentially biased) results" (10:228).

No simple solution to this problem seems to exist, so conclusions that depend on a seemingly large fail-safe number must be viewed with considerable skepticism.

A GRAPHICAL APPROACH TO RESEARCH SYNTHESIS

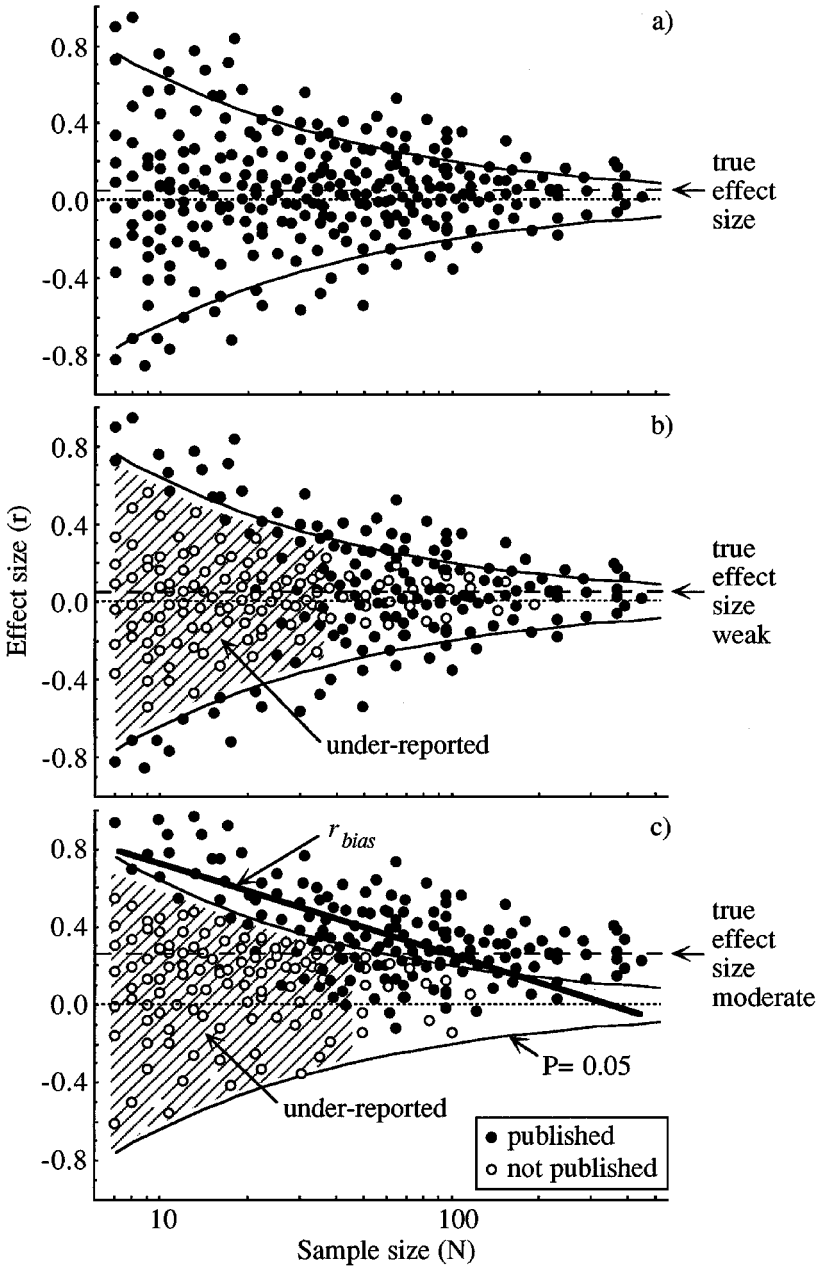
The Value of a Graphical Approach

The greatest insurance against being misled by "the pernicious influence of the modern tendency to deify the statistical significance test" (39) is appropriate graphical presentation of research results. If patterns are not apparent in simple graphical form, one may legitimately wonder about their biological significance (67).

Light & Pillemer (64) introduced a particularly attractive graphical approach to the study of selective reporting: the funnel graph (Figure 1; a scatterplot of effect size as a function of sample size). Like many graphical approaches (4), funnel graphs offer several advantages. First, they allow readers to judge for themselves how well behaved the data are: Do published results converge on some average value with increasing sample size? Are data approximately normally distributed so that a mean and standard error are appropriate statistical descriptors? Second, they allow differences between groups to be judged more easily: Are putative differences between groups of interest apparent to the eye or do they depend upon statistical wizardry (4)? Third, a funnel graph also provides a powerful exploratory tool for determining whether statistically significant heterogeneity, due to contrasts of biological or methodological interest, may have been confounded by selective reporting (80): Are studies reporting larger effects disproportionately based on smaller sample sizes?

Several meta-analyses have incorporated funnel graphs to show how effect sizes converged toward an average with increasing sample size (e.g., 6, 28, 45), but rather few have used them to test for selective reporting (32, 80, 114). Why such graphical presentations are not more widely used is hard to understand, except perhaps because, as Magnusson (67:148) wryly reflects, "[scatterplots] are not very scientific. After all, anyone, even a nonscientist, could interpret them."

Figure 1 Hypothetical funnel graphs (64)—effect size as a function of sample size (log scale)—as modified in Ref. 80. (*a*) Expected pattern of purely sample size–dependent variation in effect sizes. (*b*) The impact of selective reporting when the true effect size is weak (the classical funnel pattern). (*c*) The impact of selective reporting when the true effect size is moderate (one side of the funnel is missing, and average effect size now depends on sample size). Shaded areas and open circles indicate areas of a reduced likelihood of publication due to selective reporting. Dotted lines indicate the null hypothesis, long-dashed lines indicate overall weighted mean, and curved lines are significance levels for correlation coefficients ($P = 0.05$) from Table R of Rohlf & Sokal (89). r_{bias} refers to the correlation—sometimes significant statistically—between effect size and sample size (80).


a)
b)
c)
0.8
0.4
0.0
-0.4
-0.8
Effect size (r)
true effect size
true effect size weak
true effect size moderate
under-reported
r_{bias}
P= 0.05
published
not published
10
100
Sample size (N)

Funnel Graphs Showing No Bias

A scatterplot of effect size versus sample size (Figure 1*a*) should exhibit three predictable characteristics if results have not been influenced by selective reporting (80): (*a*) The variance of effect sizes should increase as sample sizes decrease; (*b*) the distribution of effect sizes should be normal for all sample sizes; and (*c*) the mean effect size should be independent of sample size. Any departures from these three criteria suggest nonrandom reporting of results (see Interpreting Apparent Bias in Funnel Graphs below for one alternative explanation).

Where all results come from a single study, and where sample sizes vary considerably, effect sizes should vary as expected in the absence of selective reporting (Figure 1a), unless authors have somehow censored or introduced bias into their own data. Two examples illustrate the expected pattern of purely sample size–dependent variation. In one (Figure 2*a*), the author concluded that no evidence existed for a biased sex ratio in European sparrowhawks (76). In the other (Figure 2*b*), the scatter clearly converged on a value greater than zero, indicating strong statistical support for assortative mating in water striders (6). In both cases, the data behaved as expected in the absence of selective reporting: (*a*) The variance increased with decreasing sample size, (*b*) the data were approximately normally distributed at all sample sizes, and (*c*) the mean effect size did not depend on sample size.

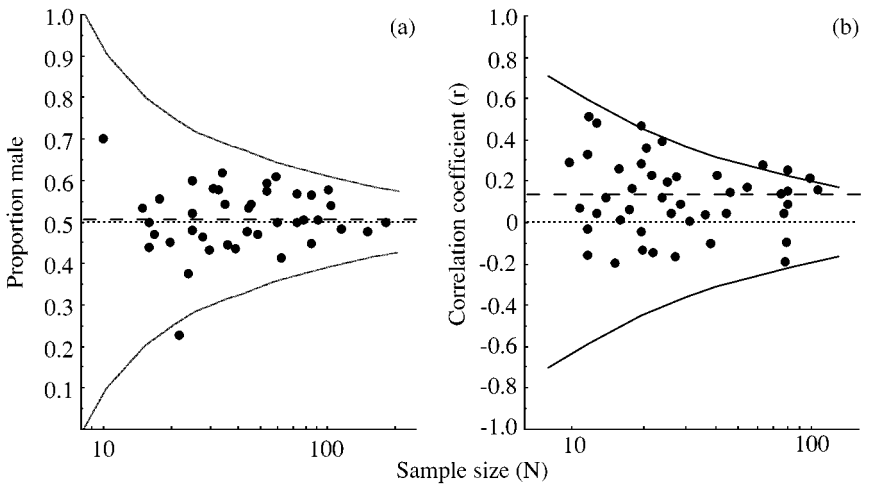


Figure 2 Two examples of within-study variation: (*a*) sex ratio variation among samples of nestlings of the European sparrowhawk (76) and (*b*) estimate of degree of assortative mating among populations of water striders (6). Dotted lines, long-dashed lines, and curved lines in (*b*) as in Figure 1. Curved lines in (*a*) are binomial significance levels ($P = 0.05$, two tailed) from Table Q of Rohlf & Sokal (89). In (*a*) the scatter converges on the null hypothesis of no sex-ratio bias with increasing sample size, whereas in (*b*) the scatter converges on a weighted mean (long-dashed line) that is different from the null hypothesis (zero).

Two Kinds of Bias Revealed by Funnel Graphs

When combined with statistical significance thresholds, a funnel graph may reveal two different manifestations of selective reporting. First, an underreporting of nonsignificant results based on small sample sizes, where the true underlying effect size is weak, yields a hole in the funnel and therefore a departure from normality (open circles and shaded area, Figure 1*b*). Second, an underreporting of nonsignificant results based on small sample sizes, where the true underlying effect size is moderate, yields a dependence of effect size on sample size (r_{bias} in 80), because results toward one side of the funnel are less likely to reach the threshold of statistical significance (open circles and shaded area, Figure 1*c*).

Interpreting Apparent Bias in Funnel Graphs

Those unfamiliar with the phrases "selective reporting" and "publication bias" might think they imply improper behavior by authors, but such an inference is unwarranted. First, because few scientists publish the results of all studies they undertake, some will inevitably lose interest if a result is not significant statistically, or not clear cut, or somehow doesn't make sense. In this respect, virtually all scientists are guilty to some degree. Second, referees and editors are not inclined to accept negative results based small sample sizes.

But even patterns consistent with selective reporting can arise for legitimate reasons. For example, a significant dependence of effect size on sample size (Figure 1*c*) may also arise for completely rational reasons: Scientists often adjust sample sizes to achieve a desired level of statistical significance. Sample-sizes might be adjusted in two ways. First, if a pilot study reveals a modest effect size, then a biologist may quite legitimately elect to use smaller samples in the final study design, since larger sample sizes are not required to demonstrate the statistical significance of the effect. In other words, sample sizes may be adjusted a priori to the size of the effect via a power analysis (98:260–65). Second, statistical significance may be monitored as data are accumulating, and data collection may be stopped once significance is achieved. Both scenarios would yield a dependence of effect size on sample size (Figure 1*c*), but only if true effect sizes genuinely varied among studies. Therefore, a dependence of effect size on sample size among heterogeneous studies is not unequivocal evidence of selective reporting.

One pattern observed among studies of fluctuating asymmetry and sexual selection—where experimental studies yield larger effect sizes than do observational ones (Figure 3)—illustrates just such a problem. Experimental studies are often based on smaller samples and yield larger effects than observational ones because the investigator has more control over extraneous factors. But if many experimental studies are conducted, and only those that reach statistical significance are published, the pattern apparent in Figure 3 may be entirely an artifact of selective reporting. The conspicuous dependence of effect size on sample size (Figure 3; see also Figures 10, 11, and 12 below) strongly suggests selective reporting.

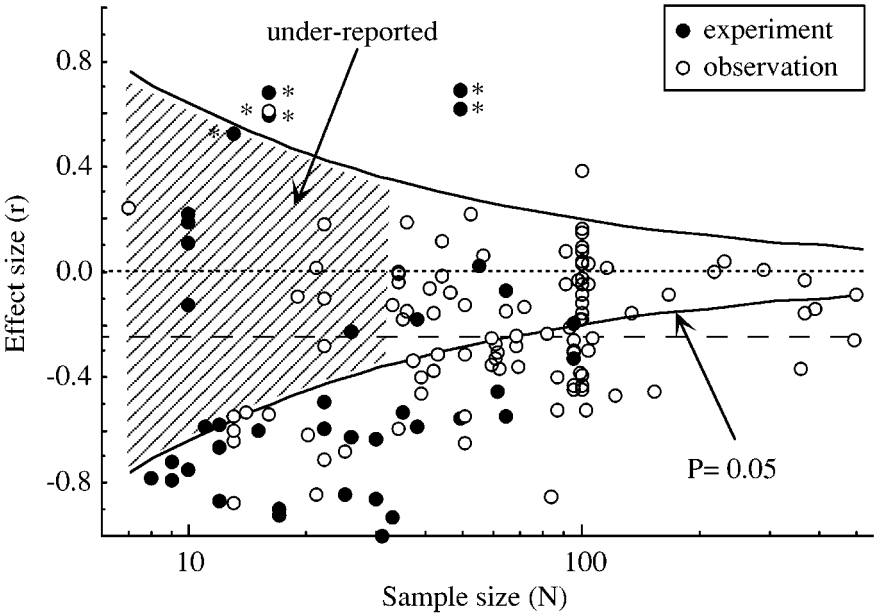


Figure 3 Effect size (correlation coefficient, *r*) as a function of sample size (log scale) for correlations between fluctuating asymmetry and individual attractiveness for experimental (●) as opposed to descriptive (○) studies (74). Dotted line, long-dashed line, curved lines, and shaded region, as in Figure 1. Modified from (80). Asterisked values were excluded from the meta-analysis by the original authors (74). The data on which this figure was based may be obtained from: http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.

The only way to distinguish between these two hypotheses for a dependence of effect size on sample size (i.e., rational planning vs selective reporting) is to conduct a formal replication of previous studies. Clearly, replicate studies should reveal approximately the same effect size regardless of sample size. If, however, the effect size is significantly lower in a replicated study based on larger sample sizes, then then the original study was biased.

Are studies based on small sample sizes necessarily a biased sample of those that were conducted? Some evidence suggests not. For example, Cappelleri et al (17) compared the results of individual large studies to those of multiple small studies in 61 meta-analyses of pregnancy and childbirth in which at least one large study was included along with a number of studies based on smaller sample sizes. Discrepancies between the single large study, and the average effect of the multiple small studies, were found in 15 of the 61 meta-analyses, but of these, 10 were explained by differences between controls ($N = 5$), differences in protocol ($N = 4$), or publication bias ($N = 1$). In the end, only one discrepancy (of 15) lacked a plausible explanation. For some human studies, where a single common effect

size is anticipated, meta-analyses of multiple smaller studies do appear to provide reliable estimates of the outcomes that would be achieved with larger sample sizes.

Limitations to Funnel Graphs Unfortunately, useful as funnel graphs may be for visualizing selective reporting (80), their utility drops dramatically as the number of studies declines (32). Funnel graphs of fewer than 30 studies may still be helpful for judging whether contrasts of biological interest may have been confounded by sample size (see Figure 3), but they will likely lack power adequate to reveal any but the most pronounced selective reporting (32).

Correcting for Bias Due To Selective Reporting

Only by exhaustively uncovering all studies conducted, whether published or not, can the biases due to selective reporting be eliminated. Such a goal would be unachievable in ecology and evolution. Both conventional (12) and Bayesian (19, 57) statistical corrections for publication bias have been proposed, but the limitations of both approaches suggest they offer no quick fix (12, 14, 43).

Ultimately, selective reporting is an inevitable element of the scientific enterprise. The challenge is to determine the extent to which it prejudices our understanding of ecological and evolutionary phenomena.

THREE EXAMPLES OF SELECTIVE REPORTING

Preamble

I have not attempted to conduct formal meta-analyses for any of the examples discussed below, nor have I systematically surveyed the literature for more recent published studies in these areas. Such endeavors lie well outside the scope of this review. In the three sections that follow, my goals are (*a*) to illustrate the value of a graphical approach to data summarization (the funnel graph; 64), and (*b*) to examine recent reviews of biological phenomena for which I could ask specific questions about what motivates biologists to publish their results. Because of the scale of this review, I was obliged to accept values reported in published summary tables as correct, and did not attempt to validate them against the original publications.

The three examples selected for scrutiny were chosen to explore different aspects of reporting patterns: (1) *sex-ratio variation*—how statistically significant associations may arise due to random sampling error and encourage detailed explanations of their putative biological significance, (2) *heritability estimates*—how a priori expectations of what results make sense theoretically influence reporting patterns, and (3) *fluctuating asymmetry and sexual selection*—how both statistical significance of results and consistency with a preferred hypothesis may influence reporting patterns.

1. Adaptive Sex Ratio Variation in Birds: Significant Associations Among Nonsignificant Samples

Studies of sex-ratio variation in birds provide a particularly illuminating example of selective reporting because on closer inspection few, if any, compelling data exist for adaptive departure from a 50:50 sex ratio in any species (20, 119), in spite of numerous published claims to the contrary. In fairness, many studies have reported little or no departure from a 50:50 sex ratio at hatching, and some have specifically drawn attention to the absence of variation greater than expected due to binomial sampling (37, 48). But new papers steadily appear claiming statistical support for adaptive sex-ratio variation (59, 94).

One attractive aspect of sex-ratio variation is explicit alternative hypotheses about which, if either, sex should predominate. Fisher (40) noted that, in the absence of confounding factors, frequency-dependent selection should promote a 50:50 sex ratio: The rarer sex will always have a relatively higher fitness. Even in the absence of selection, a 50:50 sex ratio is expected where sex is determined by the conformation of a single chromosome pair. However, local mate competition or local resource competition (reviewed in 5, 41) may promote an excess of one or the other sex.

Most departures from 50:50 sex ratio are interpreted in terms of these latter two hypotheses (e.g., see 44). However, funnel-graphs of results tabulated in two reviews suggest most, if not all, sex-ratio variation in hatchling birds does not exceed that expected due to binomial sampling variation.

Most authors who compile sex ratios seem aware that sampling error may yield spuriously significant results (e.g., 20, 44), but not all deal with it in the same way. For example, both Cockburn (21) and van Schaik (110) set an arbitrary sample size as large enough ($N = 100$ and 200, respectively), but both include studies of smaller sample sizes in their analyses if results were significant. Clearly, statistical significance in these cases outweighed the authors' a priori belief that large sample sizes were required for adequate confidence. This example also illustrates the widespread tendency to accept results based on small sample sizes if significant but to minimize or dismiss them if not significant.

Clutton-Brock's Review Following a detailed examination of published sex ratio variation in birds, Clutton-Brock (20:326–27) concluded: "Sound evidence for sex ratio variation at hatching is thus scarce. There is some evidence that the sex ratio can vary with position in clutch but trends show no consistency across populations or species. Significant relationships have been found with order of clutch ... and maternal age ... but whether these indicate that birds can vary the hatching sex ratio of their offspring in an adaptive fashion or whether they represent the small proportion of cases where the null hypothesis has been wrongly rejected by chance is as yet not certain." These conclusions echoed an earlier one by Williams (119) that the data for birds did not support any theory of adaptive sex ratio evolution.

Two lines of evidence support Clutton-Brock's (20) suspicion that sampling error alone accounted for the observed sex-ratio variation. First, of 85 separate sex-ratio estimates from 14 studies of 8 species, 11 differed significantly from parity at $0.05 > P > 0.01$ and 2 were significant at $P < 0.01$. These numbers do not differ significantly from those expected due to sampling error ($P = 0.116$; Chi-square test with correction for continuity), although slightly more significant studies appear to have been reported ($P = 0.035$) when they are divided into two groups, those significant at $P < 0.05$ and those not. Second, a funnel-graph of these data clearly reveals the pattern expected for purely random, sample size-dependent variation in sex ratio (Figure 4*a*).

Furthermore, among studies where the title of the original paper stated or implied a biased sex ratio (i.e., where the authors wished to emphasize a significant departure from parity), a closer examination suggests that statistical significance arose due purely to sampling error. First, the sample sizes of these studies were significantly smaller than those of the other studies in Clutton-Brock's review

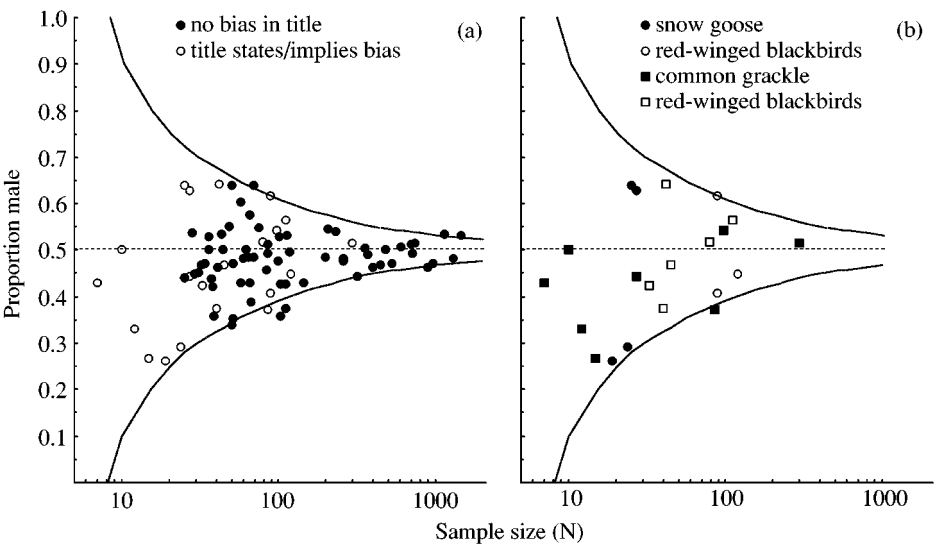


Figure 4 Variation in the sex ratio of samples of birds as a function of sample size (log scale) (all data from Ref. 20). Single species are represented by multiple points and, in some cases, data for single species come from multiple papers. Dotted lines indicate the null hypothesis, and curved lines indicate binomial significance levels ($P = 0.05$, two tailed) from Table Q of Rohlf & Sokal (89). (*a*) samples differentiated by title of paper (does the title state or imply a biased sex ratio was observed?), (*b*) samples for single species exclusively from the four papers with a title stating or implying a biased sex-ratio [●, position in egg-sequence (3); ○, maternal age (15); □, time in laying season (115)]. The data for grackles [■ (53)] are not comparable because the sex-ratio variation tabulated by Clutton-Brock was unrelated to the main point of this paper. The data on which this figure was based may be obtained from: http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.

(<0.001, Mann-Whitney U-test; Figure 4*a*). Second, although only 2 of the 21 individual samples actually deviated significantly from a 50:50 sex ratio, 3 studies reported significant associations with sex ratio (Figure 4*b*). In other words, random variation among nonsignificant samples yielded statistically significant associations that seemed biologically interesting and could easily be interpreted in light of one or another theory of sex allocation.

Gowaty's Review In contrast to Clutton-Brock (20), Gowaty (44) concluded that small sex-ratio differences between passerine and anseriform birds were significantly correlated with sex-dependent patterns of philopatry: The sex that dispersed further was the sex that tended to be overproduced. Even though sex ratios appeared to depart significantly from parity in only 5 of the 12 passerine species, and in none of the ducks and geese (Figure 5), Gowaty noted that what deviations did exist (whether significant or not) tended toward excess males in ducks and geese (5 of 6 species) and excess females in passerine birds (11 of 12 species) (Figure 5). Gowaty concluded that the consistent directions of these differences were too improbable to be due to chance ($P = 0.004$ at the level of species, and $P = 0.036$ at the level of families; Fisher's exact test), and that "the lack of statistically significant differences from a 50:50 sex ratio may have obscured biologically interesting phenomena associated with sex ratio variation in birds" (42:272).

However, a closer inspection of the data and reasoning raises doubts about these conclusions. First, a funnel graph once again revealed a pattern consistent with simple sampling variation (Figure 5). Second, species in two other orders also exhibited sex-biased philopatry, but in these two orders, either the overproduced sex

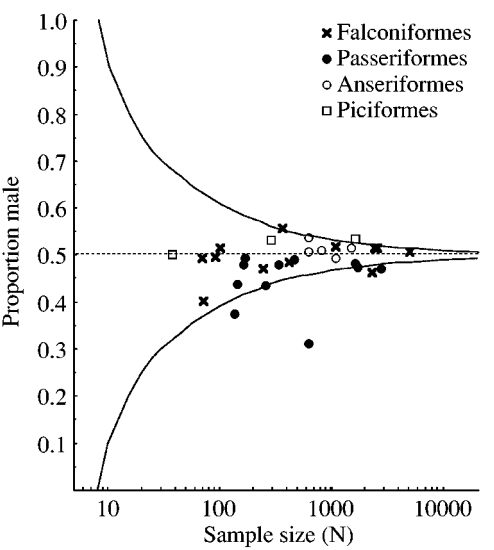


Figure 5 Variation in the sex ratio of nestling or fledgling bird species within each of four orders as a function of log(sample size) (all data from Ref. 44). Dotted line and curved lines as in Figure 4.

TABLE 1 Number of species or families exhibiting an excess of male or female offspring (from Ref. 44)[a]

Order	Philopatric Sex	Number of Species (Families) with Excess Males	Number of Species (Families) with Excess Females	Number of Species (Families) with 50:50 Ratio
Anseriformes	female	5(2)	1(0)	
Passeriformes	male	1(0)	11(6)	
Falconiformes	male	6(2)	6(0)	
Piciformes	male	2(1)	0(0)	1(0)
Male-philopatric pooled		9(3)	17(6)	1(0)

[a]Sex ratios for both species and families were computed using weighted averages (counts of males and females were pooled separately across samples and a new ratio computed from these pooled values).

was uncorrelated with the dispersing sex (Falconiformes) or the overproduced sex dispersed less (two of three woodpecker species) at the level of species or families (Table 1). When all male-philopatric species were pooled, the dependence of sex ratio on philopatry ceased to be significant ($P = 0.09$, contingency table analysis corrected for continuity). Third, sexual size dimorphism—the presumed raison d'etre for the departures from 50:50 sex ratio (44)—is negligible in both passerines and ducks and geese. So the statistical significance of the apparent correlation between sex ratio and philopatry in birds depended on the orders examined.

Conclusion When the bird sex-ratio data are viewed as a whole (Figures 4 and 5), two patterns emerge. First, even though authors may have drawn attention to results that were statistically significant, they also reported results that were not significant. As a consequence, statistically significant samples were only weakly overreported. Second, the pooled data support rather strongly a tightly constrained 50:50 sex ratio, subject to little more than sampling error (Figure 4*a*), as both Williams (119) and Clutton-Brock (20) surmised might be true. The biological significance of the many studies reporting statistically significant departures from parity (44, 59, 94) is therefore questionable.

The advent of modern technologies that allow sexing prior to or shortly after hatching using blood samples (flow cytometry, DNA profiles, microsatellites) has inspired additional studies of sex ratio variation in birds (94) because such techniques can minimize or eliminate effects of sex-biased mortality after hatching. Although some seemingly compelling examples are mentioned (e.g., 58), a more detailed funnel-graph analysis, or explicit replication of some of these results by others, would be required to reject convincingly the possibility that they too are simply examples of selective reporting.

In view of the many reports of statistically significant sex-ratio variation in birds that appear to have arisen simply due to sampling error, I encourage all of those tempted to offer biological explanations for such patterns to restrain their

enthusiasm until some of the more striking claims have been replicated independently by others. Sex ratios are too easy to measure incidentally as part of another study, and therefore they would seem particularly prone to selective reporting. Sex ratios obtained with modern techniques (94) may yield values less confounded by differential mortality, but they do not avoid the fundamental problem: Departures from 50:50 require rather large samples to detect reliably (Figures 4 and 5).

2. Heritability: Impact of Theoretical Preconceptions

That theoretical preconceptions alone have an impact on probability of publication is perhaps most clear cut among studies of heritability. Narrow-sense heritability describes the degree to which offspring resemble their parents on average (35). Few biologists would expect offspring to exhibit consistently more extreme phenotypes than their parents (heritability >1.0), or for offspring to deviate consistently from the population mean in the opposite direction from their parents (a negative heritability). But sampling error should yield such heritabilities occasionally, just due to chance (66). In addition, as sample size decreases, the likelihood of a negative or extreme positive heritability increases substantially.

Published estimates of narrow-sense heritabilities (116) reveal what appears to be a clear example of selective reporting (Figure 6). In part, this is influenced by statistical significance: Nonsignificant studies are underrepresented at small sample sizes and average heritability decreases significantly with increasing sample size ($P < 0.001$, Figure 6). More seriously, even at small sample sizes ($N < 50$), where heritabilities are estimated with lower confidence, negative heritabilities are virtually nonexistent, even though several positive estimates exceed the theoretical maximum (Figure 6). At face value, authors appear more comfortable with super-heritability ($h^2 > 1.0$) than with negative heritability ($h^2 < 0.0$).

Studies of the heritability of fluctuating asymmetry reveal the same pattern even more dramatically (Figure 7). The weighted mean heritability is closer to zero than in the previous example (compare to Figure 6), but again, virtually no heritability estimates were less than zero. The absence of any but the slightest negative heritability estimates in both examples (Figures 6 and 7) reveals rather clearly that theoretical expectations influence the likelihood of publication, quite independent of any effects of statistical significance.

Where are all the missing negative heritability estimates? Most likely, they reside in filing cabinets because they made no sense theoretically.

This result is troubling because it implies that even carefully conducted meta-analyses could yield statistical support for a preconception, rather than a genuine biological phenomenon, if the theoretical grounds for that preconception are strongly held (but see Figure 14). For example, consider the following thought experiment. Assume that the true heritability of a particular trait is zero. If 100 biologists independently estimate this heritability, sampling error dictates that half the observations should be negative and half positive. If those biologists who obtain negative heritabilities discard their results, or set them to zero as advised by some

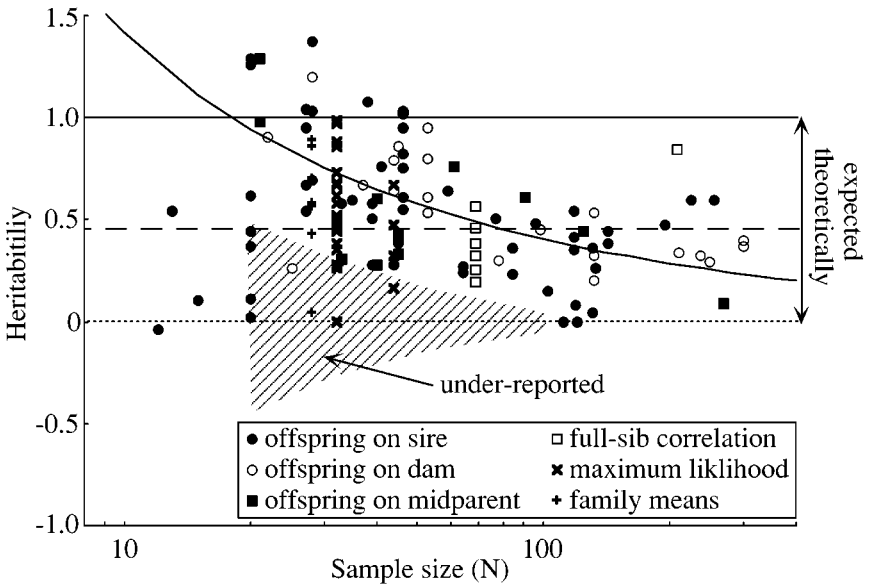


Figure 6 Published heritability estimates (h^2) from field studies as a function of sample size and method of estimation (116). Sample sizes (number of families) were obtained from the original studies. The dotted line indicates the null hypothesis, the long-dashed line indicates the weighted mean heritability across all studies (0.452), and the curved line indicates the upper 95% confidence interval (= 2SE) about zero for heritability estimates according to Falconer (35:166) for parent-offspring regressions based on one offspring per parent ($2/\sqrt{(N-2)}$). This confidence interval is not appropriate for all the methods indicated but does indicate how the significance level varies with sample size. Over all samples, heritability estimates decreased significantly with increasing sample size (Spearman ρ corrected for ties $= -0.33$, $P < 0.001$). The shaded region indicates where observations are expected due to sampling variation but have been underreported. The data on which this figure was based may be obtained from: http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.

practitioners (35), what is the net result? A meta-analysis would likely yield strong statistical support for a positive heritability that is entirely an artifact of our preconceptions. One can only wonder at how many published heritability estimates have been exaggerated by this preconception.

3. Fluctuating Asymmetry and Fitness: Anatomy of a Bandwagon

Fluctuating asymmetry (FA; small, random departures from perfect symmetry; 65) offers an intuitively appealing measure of developmental precision (the degree to which the right and left sides of a bilaterally symmetrical organism depart from perfect symmetry due to the cumulative effects of developmental noise) (79, 82). It is appealing because of the apparent ease with which it may be measured,

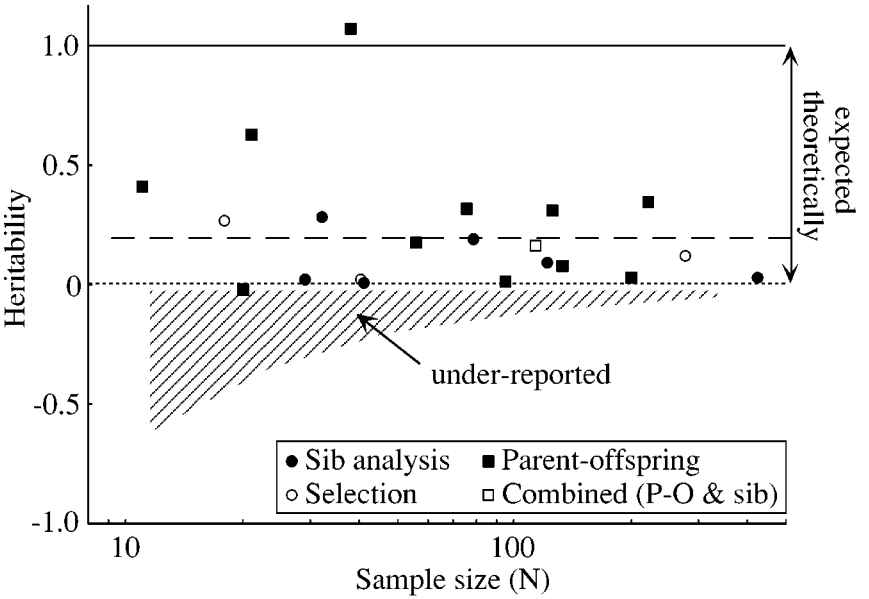


Figure 7 Published estimates of the heritability of fluctuating asymmetry (h^2) for various taxa and methods of estimation as tabulated on the internet site to accompany Ref. 73. Sample sizes cannot be interpreted literally here because very different methods were used in different studies; only the distribution of heritability estimates is relevant. The dotted line indicates the null hypothesis and the long-dashed line indicates the weighted mean heritability across all studies ($h^2 = 0.154$). The shaded region indicates where observations are expected due to sampling variation but have been under-reported. The data on which this figure was based may be obtained from: http://www1.oup.co.uk/MS-asymmetry.

and because of the seemingly sound theoretical grounds for believing that subtle departures from symmetry really should reveal something about underlying developmental stability (69, 82, 111).

Over the last 10 years, interest in FA has increased more than 10-fold (Figure 8). Many biologists have rushed to apply this approach to a variety of questions, since developmental precision is thought (*a*) to be reduced by environmental or genetic stress (see 122 and references therein) and by lowered heterozygosity (113), and (*b*) to correlate negatively with measures of individual fitness (72)—including growth, fecundity, and survival—and with measures of individual attractiveness in studies of sexual selection (74, 104, 105).

Three lines of evidence, however, suggest that selective reporting has greatly exaggerated the apparent strength and generality of one association: the correlation between individual asymmetry and measures of individual quality, fitness, or attractiveness. This evidence includes (*a*) the absence of parallel asymmetry variation among individuals, (*b*) theoretical demonstrations of the limited

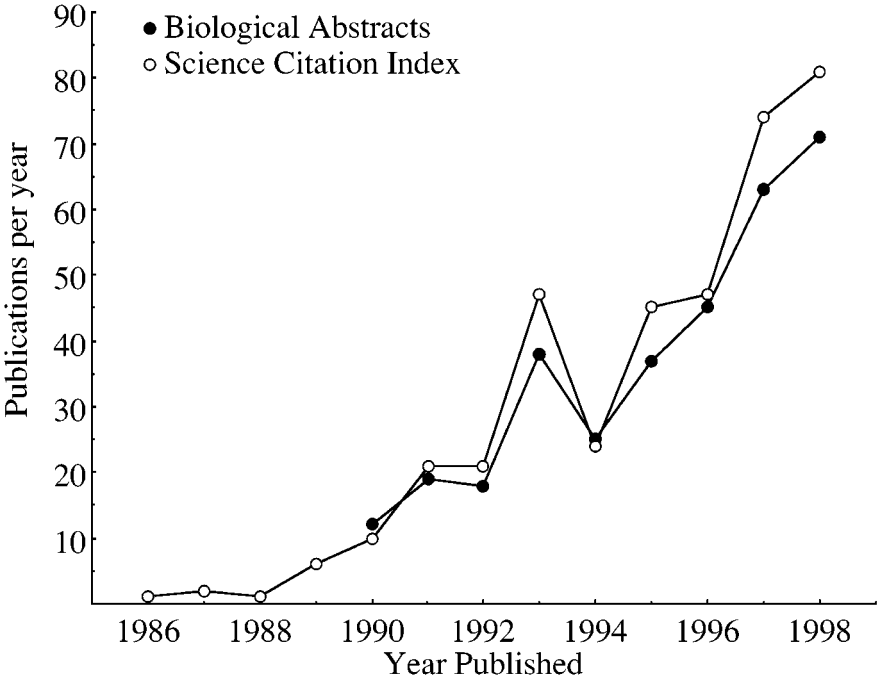


Figure 8 Number of citations per year that include "fluctuating asymmetry" in the title, abstract, or keywords in the electronic database of Biological Abstracts (●), or cite the 1986 review by Palmer and Strobeck (82), as determined from Science Citation Index (○).

statistical power of FA as a measure of underlying developmental instability, and (*c*) patterns of selective reporting among studies of FA and sexual selection that are not observed in other studies of FA variation.

Absence of Parallel Asymmetry Variation Among Individuals If subtle deviation from symmetry is a reliable indicator of underlying developmental stability in an individual, as claimed repeatedly in the FA literature (reviewed in 73), then deviations from symmetry in one trait should correlate with deviations from symmetry in other traits on that same individual. The virtual absence of parallel asymmetry variation among individuals for morphological traits, however, raises serious doubts about whether asymmetries of individual traits should be correlated with any other phenomena of biological interest (52, 78).

An early review of FA variation (82) noted that while asymmetries might be correlated among populations (a population more asymmetrical for one trait was typically more asymmetrical for others—the "population asymmetry parameter" 99), they were rarely correlated among individuals within populations (the "individual asymmetry parameter"; 62). And more recently, an extensive review of

FA variation in many traits in a variety of organisms reached the same conclusion (73:53–55).

Limited Statistical Power The limited ability to detect parallel asymmetry variation among individuals arises from two complementary causes: sampling error and measurement error.

Not until 1994 was a simple statistical explanation advanced for the rarity of parallel asymmetry variation among individuals (78:360): The absolute value of the deviation from symmetry in a single trait of an individual estimates the underlying right-left variance for that trait with only one degree of freedom (see also 107, 117, 118). In other words, the absolute difference between sides in one trait of an individual estimates the underlying developmental instability (the potential right-left variance) of that trait with only one degree of freedom. As most biologists know, the ability to detect differences among means is limited if they are estimated with only one degree of freedom each, and the ability to detect differences among variances is considerably more limited (97).

Simulations confirm this expectation (107). In the absence of measurement error, the observed correlation between asymmetries in two traits was only 0.287 ($r^2 = 0.082$), even where the underlying instability variance differed by 16-fold (Figure 9*a*). So only about 8% of the asymmetry variation in one trait can be predicted by asymmetry variation in a second on the same individual, even under ideal conditions: sizeable variation in underlying developmental stability among individuals and no measurement error.

Measurement error further reduces the expected correlation between asymmetries. It can form a sizeable fraction of the between-sides variation because FA variation is often on the order of 1% of trait size (38, 79) and few biologists measure traits to a precision much greater than 1%. A high percent measurement error significantly reduces the strength of asymmetry correlations. For example, only 3% of the variation in $|R - L|$ in one trait is explained by variation in $|R - L|$ of a second in the same population when measurement error is half of the between-sides variance (Figure 9). In addition, direct evidence suggests that as the repeatability of asymmetry measures increased in published studies, so did the strength of asymmetry correlations (109).

Undaunted, those who believe in the predictive value of FA have invoked this low statistical power to their advantage. For example, Gangestad & Thornhill (42) argued forcefully that the theoretical maximum correlation between individual asymmetry and attractiveness should be -0.27 and therefore that the observed weighted mean effect size of -0.22 implies a high proportion of variation in attractiveness can be attributed to variation in underlying developmental instability (105). That they remained unperturbed by the majority of effect sizes that exceeded this putative theoretical maximum in their own tabulation is a testament to the strength of their convictions.

Clearly, on purely statistical grounds, the rarity of parallel asymmetry variation among individual organisms is not surprising. Even if a true correlation between

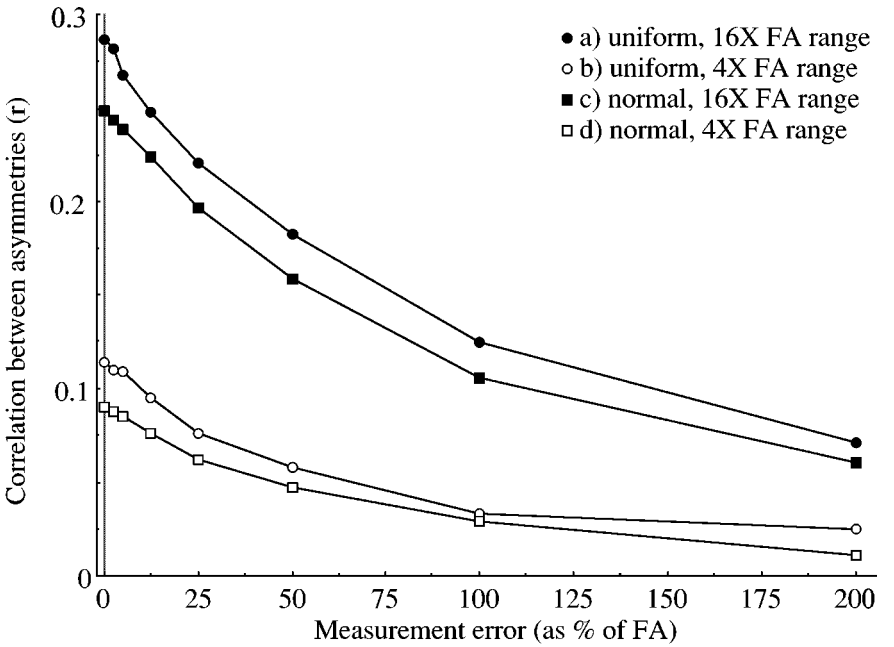


Figure 9 Effect of measurement error on the strength of the correlation (r) between absolute deviations from symmetry (FA) in two traits among individuals in a single sample; kindly simulated by S Van Dongen using the same model as in Ref. (107), but for 100,000 replications. Each population consisted of a mixture of individuals exhibiting three different levels of underlying instability variance: var($R-L$) = FA = 1/x, 1, x. Two populations were simulated, as were two different distributions of FA variation: (●) x = 4, proportions of all three FA levels equal; (○) x = 2, proportions of all three FA levels equal; (■) x = 4, proportions of FA levels 1:2:1; (□) x = 2, proportions of FA levels 1:2:1. The measurement error variance var(M_1-M_2) is expressed as a percent of the median instability variance (i.e., a value of 100 means the variance of replicate measurements equals the median FA variance between sides).

FA and some trait of interest is 0.2 (close to the putative theoretical maximum), the statistical power of correlation coefficients is low for routine sample sizes (83): With a sample size of N = 40, a significant correlation ($P \leq 0.05$) would be detected only about 20% of the time (power = 0.2), and even with a sample size of N = 100, the power is less than 0.5. In other words, even for a true correlation close to the theoretical maximum, sampling variation should not yield a significant correlation more than 50% of the time.

What accounts for the remarkable number of statistically significant correlations reported between individual asymmetry and other features of animals such as fitness or attractiveness? As Houle (52) noted so pointedly, it is hard to understand how correlations between individual subtle asymmetries and other phenomena of interest can be so common when correlations between asymmetries

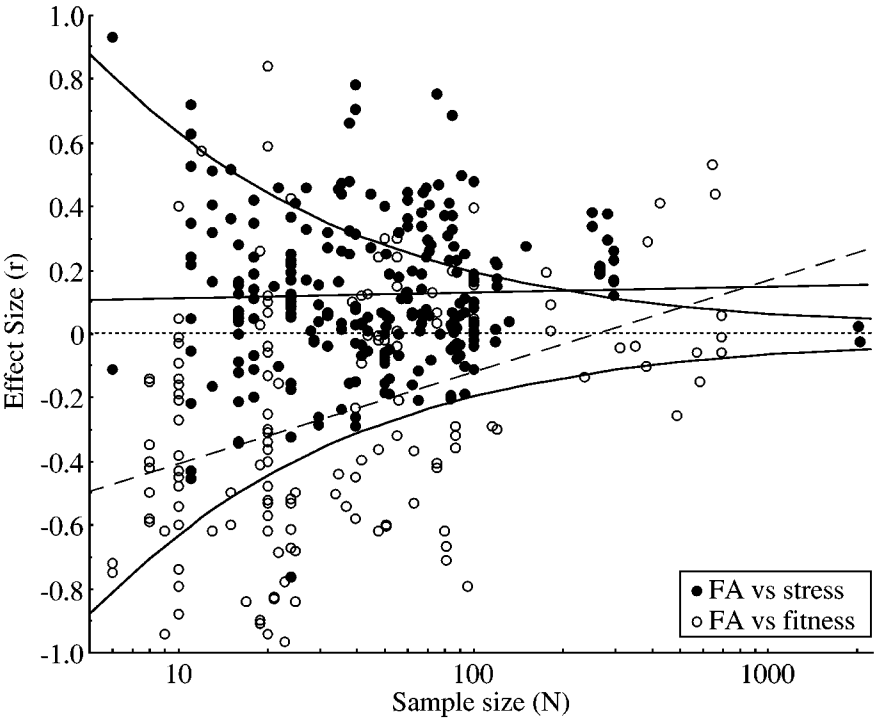


Figure 10 Effect size (correlation coefficient, *r*) as a function of sample size (log scale) for correlations between fluctuating asymmetry (FA) and stress (●) or various measures of fitness (○) as reported in the meta-analysis by Leung & Forbes (63). Dotted line and curved lines as in Figure 1. The solid line indicates the least-squares linear regression for stress studies ($Y = 0.016\ X + 0.098$, $N = 151$ cases; Spearman's $\rho = 0.031$, $P = 0.63$), and the long-dashed line indicates the regression for fitness studies ($Y = 0.285\ X - 0.691$; Spearman's $\rho = -0.39$, $P < 0.001$). Effect sizes are expected to be positive for FA-stress relations (higher stress results in higher FA), but negative for FA-fitness relations (higher FA is associated with lower fitness). Measures of fitness included traits like body size, mating success, dominance, secondary-sexual trait size, survival, condition, and growth. The data on which this figure was based may be obtained from http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.

in the same individuals are so rare. Selective reporting seems a likely explanation.

Direct Evidence of Selective Reporting Claims that seem highly improbable (81) certainly raise the possibility of selective reporting (see http://www.biology.ualberta.ca/palmer.hp/asym/Curiosities/Curiosities.htm for some examples). Furthermore, evidence from three meta-analyses strongly suggests that selective reporting of associations between FA and various measures of individual fitness may be a serious problem.

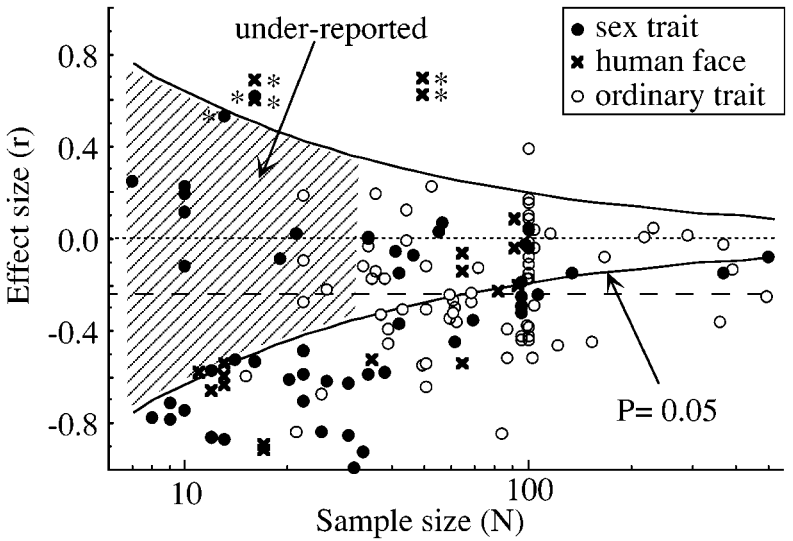


Figure 11 Effect size (correlation coefficient, r) as a function of sample size (log scale) for correlations between fluctuating asymmetry and individual attractiveness for sex traits (●), human faces (**x**), and ordinary traits (○) (as tabulated in Ref. 74, figure modified from Ref. 80). Dotted line, long-dashed line, curved lines and shaded region, as in Figure 1. Asterisked values were excluded from the original analysis (74) on methodological grounds. The data on which this figure was based may be obtained from http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.

The dependence of effect size on sample size, r_{bias} (80), was highly significant statistically for signaling traits (sex + human face; Spearman's $\rho = 0.38$, P = 0.002, N = 67) but not for ordinary or nonsignaling traits (Spearman's $\rho = 0.16$, P = 0.19, N = 73). The same patterns were apparent when experimental studies were excluded: r_{bias} was highly significant for signaling traits ($\rho = 0.44$, P = 0.009, N = 36), but not for ordinary traits ($\rho = 0.12$, P = 0.34, N = 69).

First, Leung & Forbes (63:400), in the earliest meta-analysis of FA variation, concluded that overall correlations between FA and stress, and between FA and various fitness measures, were "non-spurious" but "fairly weak, and highly heterogeneous." A closer examination (Figure 10) reveals that FA-stress correlations were largely independent of sample size ($r_{bias} = 0.031$; P = 0.63) and remained more or less centered on the overall weighted mean of $r = 0.17$. A similar nonsignificant r_{bias} was found in a meta-analysis of FA-heterozygosity correlations (113). FA-fitness correlations, however, revealed a different pattern: As sample size increased, effect size decreased significantly ($r_{bias} = -0.39$, P < 0.001). Fully 16% of the overall variation in effect sizes could be attributed to variation in sample size. Average effect sizes appeared moderate (0.3–0.5) (7) when based on sample sizes less than 20, but for sample sizes greater than 50, they were weak (<0.2).

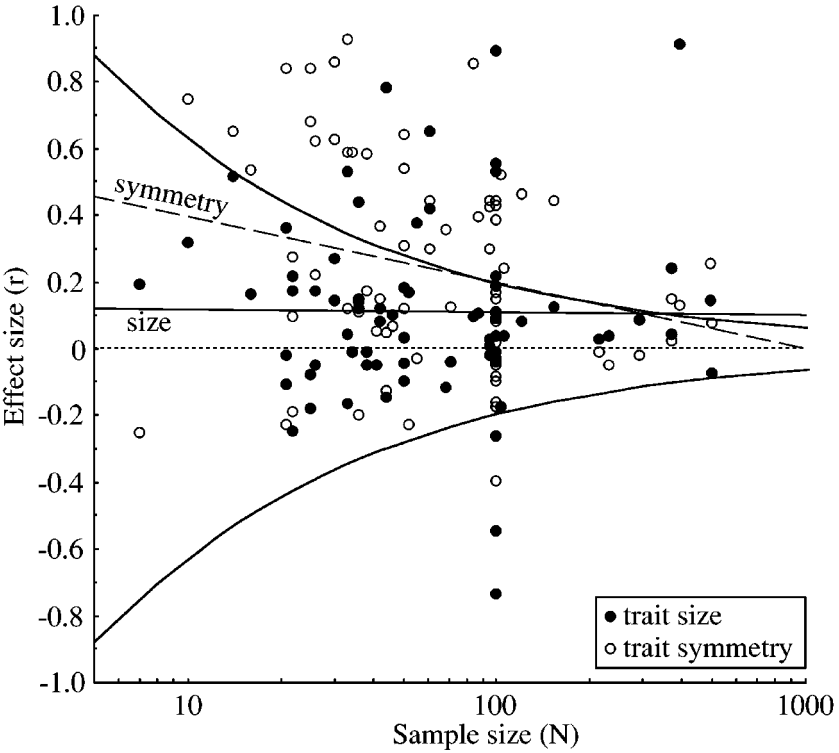


Figure 12 Effect size (correlation coefficient, *r*) as a function of sample size (log scale) for correlations between trait size and individual attractiveness (●) and correlations between symmetry and individual attractiveness (○) (as tabulated in Ref. 104). Dotted line and curved lines as in Figure 1. The solid line indicates the least-squares linear regression for trait size ($Y = -0.010\ X + 0.130$, $N = 73$ cases; Spearman's $\rho = -0.054$, $P = 0.65$), and the long-dashed line indicates the regression for fitness studies ($Y = -0.204\ X + 0.609$, $N = 73$ cases; Spearman's $\rho = -0.260$, $P = 0.027$). Effect sizes are expected to be positive for both relations (larger or more symmetrical traits are more attractive). The trait symmetry data are the same as those in Figure 11 (74), but were limited to studies where both trait size and trait symmetry were examined simultaneously. In addition, the sign of the asymmetry effects was reversed to permit direct comparisons between effects of symmetry and size.

Such a relationship renders statistical summaries of weighted-mean effect sizes virtually meaningless, since average effect size clearly depends on sample size.

Second, a more restricted meta-analysis that specifically examined correlations between asymmetry and attractiveness (74), rather than between asymmetry and a variety of fitness measures, revealed an even stronger suggestion of selective reporting (80): (*a*) The threshold of statistical significance ($P = 0.05$) rather sharply defined the upper boundary to a cluster of published associations based on sample

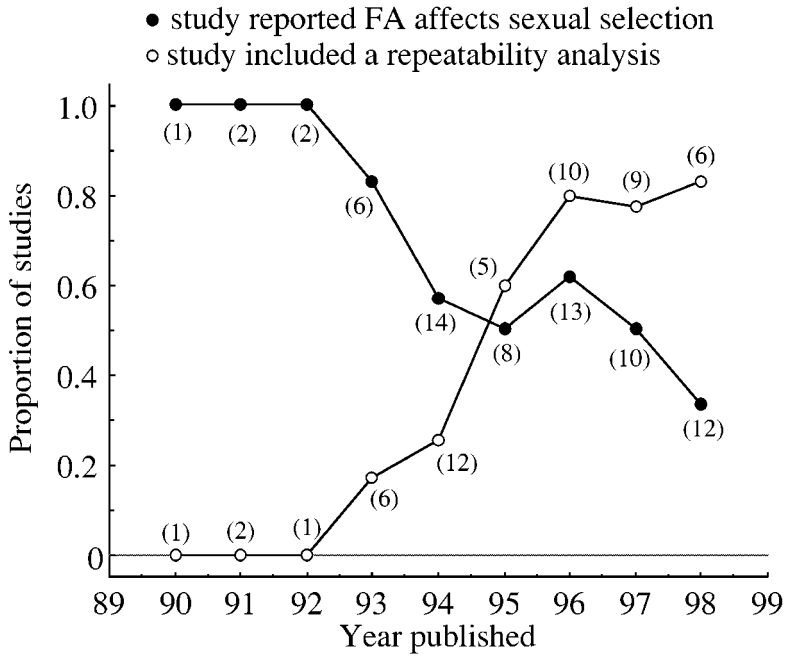


Figure 13 Change over time in the proportion of studies that reported an association between fluctuating asymmetry (FA) and sexual selection and the percent of studies that conducted a test for the repeatability or reliability of the FA signal relative to measurement error (from Ref. 96). Numbers in parentheses indicate number of studies.

sizes less than 20 (lower left portion of Figure 11); (*b*) overall effect size declined significantly with increasing sample size (r_{bias} = Spearman $\rho = 0.39$, $P < 0.001$, $N = 140$; 'included' samples only); (*c*) r_{bias} was highly significant for signaling traits ($P = 0.002$) but not for ordinary or nonsignaling traits ($P = 0.19$). The same patterns were apparent when experimental studies were excluded (r_{bias} was highly significant for signaling traits, $P = 0.009$, but not for ordinary traits, $P = 0.34$), so the disproportionate number of experimental studies at small sample size (Figure 3) was not the cause of the original pattern. Perhaps most seriously of all, r_{bias} was highly significant for signaling traits but not for ordinary traits among the studies conducted by Møller and Thornhill themselves (see Figure 3 of Ref. 80).

Third, another meta-analysis (104) examined cases in which the effects of both trait size and trait symmetry on mating success or attractiveness were studied concurrently. As above (Figure 10), r_{bias} was significant for symmetry variation ($P = 0.027$) but not for variation in trait size ($P = 0.65$) (Figure 12).

Finally, some surprising evidence suggests that many of the initial reports trumpeting the role of FA in sexual selection were actually spurious results that arose from high levels of measurement error coupled with selective reporting. Many early

studies did not test whether the putative FA variation being reported exceeded that due simply to measurement error, even though measurement error yields bilateral variation indistinguishable from true FA and was clearly recognized as a serious problem for FA studies (82). As more studies tested for the significance of FA variation relative to measurement error, fewer and fewer reported significant associations with FA (Figure 13). This change over time remains striking, even if the small number of studies prior to 1993 are ignored.

All of the above evidence suggests that studies of FA and individual fitness or attractiveness have been seriously confounded by selective reporting, particularly since a significant r_{bias} was absent among studies of FA and stress (63), FA and heterozygosity (113), and trait size and attractiveness (104). In the end, studies of FA and sexual selection, or FA and individual fitness, will likely reveal more about the sociology of science than about biology.

Alternative Hypotheses to Selective Reporting: Disproof by Replication

The preceding three examples suggest selective reporting may have promoted dubious biological conclusions. This hypothesis is open to disproof. If genuinely replicated studies reveal effect sizes of magnitude and direction similar to those of the original results, then the hypothesis of selective reporting for these cases is rejected.

Sex Ratio Variation Sex-ratio variation in birds, in two formal reviews, appears not to exceed that expected due to sampling error. No doubt this will trouble many who believe otherwise. The hypothesis of purely random variation may be rejected readily by one or two truly replicated studies of published claims of striking departures from parity. For example, if sample sizes exceeding 200 revealed comparable departures from a 50:50 sex ratio in different eggs in the laying sequence, as reported for snow geese based on sample sizes less than 30 (3), then the hypothesis of selective reporting can be rejected. In fact, this replication has already been conducted. Cooke & Harmsen (22), based on a more detailed study (though not on much larger sample sizes), found no statistical support for a dependence of sex-ratio on laying sequence. Seasonal sex-ratio variation in red-winged blackbirds is similarly suspect because two independent studies yielded contradictory results (37, 115). Similarly, if sex ratio of offspring varied by similar amounts among mothers of different ages in red-winged blackbirds based on sample sizes of 300 instead of about 100 (15), then the hypothesis of selective reporting can be rejected.

The only way to determine whether the few significant associations reported by Clutton-Brock (20) and Gowaty (44) were due to chance would be to repeat some of the original studies.

Heritability Negative heritability estimates appear to be greatly underreported in the literature (Figure 6). Among studies of the heritability of FA prior to 1998, such underreporting appears quite pronounced (Figure 7). The claim that FA variation

is significantly heritable (73), and the use of that claim to buttress other preferred hypotheses (73), thus seems open to question.

One recent study of the heritability of FA variation reveals just how variable heritability estimates may be (121). Not only does heritability vary by several fold among traits and populations, but as many estimates are negative as are positive (Figure 14). Furthermore, some of the negative estimates are more extreme than the most extreme positive ones. Woods et al (121) are to be commended for reporting such results. Even though negative heritability estimates may make no sense theoretically, they should arise due to sampling error (66). If such negative heritabilities were excluded from prior studies, then the apparent average significant heritability suggested by Figure 7 (73) and elsewhere (108) may be largely or entirely an artifact of selective reporting.

One or two formal replications of earlier studies, particularly of those that reported highly significant heritabilities of FA variation (e.g., $h^2 = 0.63$ in stickleback lateral plate numbers) (46) and ($h^2 = 1.072$ in scorpionfly forewings)

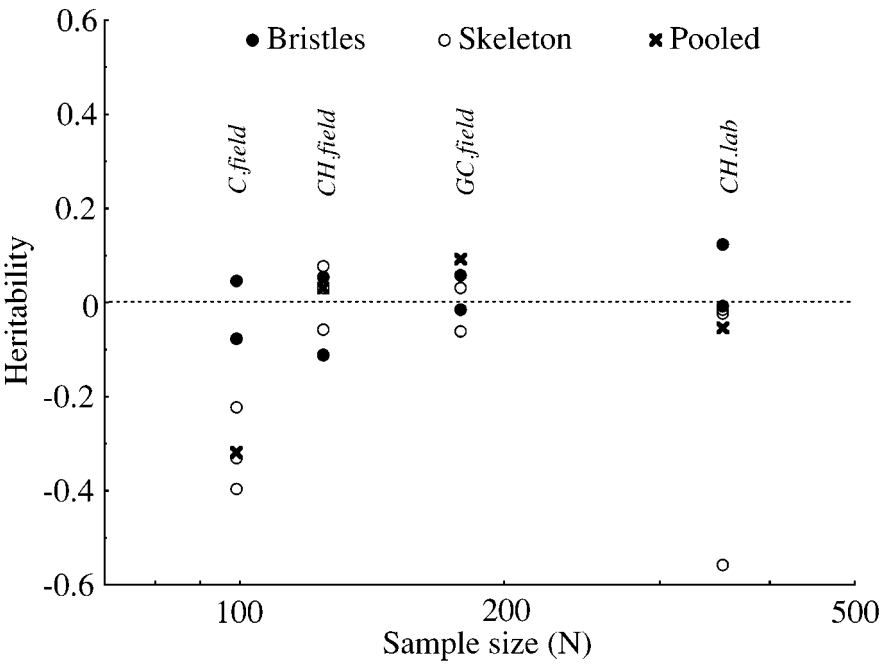


Figure 14 Estimates of heritabilities (parent-offspring regressions) in *Drosophila melanogaster* (121). Estimates were obtained for two bristle and three skeletal traits in three field and one laboratory population. The dotted line indicates the null hypothesis. N refers to the number of families in which "one to two" female offspring were measured per female parent. Abbreviations: C, Cairns; CH, Cherry Hill; GC, Gold Coast. The data on which this figure was based may be obtained from http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.

(106) would provide a far more robust test of the true heritability of FA variation than additional quasireplications in other taxa or traits. If these high heritabilities could be repeated, then, again, the hypothesis that the overall significant heritability of FA (Figure 7) is due to selective reporting would be rejected.

Fluctuating Asymmetry and Sexual Selection Remarkably, in spite of the huge impact Møller's (71) original study had on interest in asymmetry as a measure of individual attractiveness, I am unaware of any independent published replication of the claim that female barn swallows (*Hirundo rustica*) prefer males with outer tail feathers that are more symmetrical, or of any other claims for correlations between asymmetry and individual attributes in barn swallows or other organisms. In view of the large number of studies of correlations between individual asymmetry and fitness/attractiveness, the absence of replicative studies seems remarkable.

If any claims for correlations between subtle asymmetries and attributes of individuals can be repeated by others with no vested interest in a particular outcome, only then will such claims earn the respect of credibility. Until such time, the hypothesis that correlations between subtle asymmetries and attributes of individuals result predominantly from selective reporting remains a viable one.

DISCUSSION

I have found these results particularly sobering and have gained little pleasure from summarizing them. Clearly all of us—meticulous and conscientious as we may be—are guilty of selective reporting to some degree. Furthermore, those who have tried to publish negative or nonsignificant results may have been discouraged or denied by well-intentioned peers or editors in the review process. As a consequence, what gets published is inevitably not a random sample of studies that were initiated, and we cannot escape the troubling conclusion that some—perhaps many—cherished generalities are at best exaggerated in their biological significance (80) and at worst a collective illusion nurtured by strong a priori beliefs often repeated (33, 73).

Quasireplication likely exacerbates the problem of selective reporting. Authors who explicitly set out to replicate a prior study fully will presumably have a greater desire to publish their results—no matter what the outcome—than those who pursue simply affirmative quasireplication. Quasireplication is still valuable because results and hypotheses ultimately do need to be repeated in other systems to assess their generality. But if quasireplication continues to substitute for formal replication, the pernicious effects of selective reporting will do little more than reinforce the contract of error that entrenches a priori beliefs or perpetuates the unconstructive cycle of bandwagon and backlash (2, 84, 86, 96). Unfortunately, editorial policies may contribute to the problem.

The Impact of Editorial Policy

Editorial policies of many scientific journals appear to reinforce a culture in which original research is valued more highly than replicative research, and in which statistically significant results are favored over nonsignificant ones (26, 50). For example, many journals explicitly discourage replicative studies either in their advice to authors (Table 2a) or in their advice to reviewers (Table 2b). As a consequence, because truly replicative research is so difficult to publish, biologists test the validity of hypotheses or patterns by examining other species or systems.

In fairness, conspicuous examples of replicated studies are sometimes published in premier journals. For example, *Science* reported a study (88) that was unsuccessful at replicating the results of an earlier one, published in *Nature-Genetics*, claiming to have found microsatellite markers linked to human male homosexuality (54). *Science* also drew attention to the great difficulties different labs had at obtaining the same results even in a highly controlled study of mouse behavioral genetics, and concluded by saying "every result should be replicated with a new batch of mice within the same lab, and perhaps elsewhere, before it's published" (34:1599). In these cases, however, conflicting results appeared to make better press than confirmatory ones.

However, even strictly confirmatory results have been deemed of sufficient importance by *Nature* to warrant publication, so long as the subject—Neanderthal

TABLE 2a Portions of instructions to authors from various journals

Journal	Instructions to Authors
American Naturalist	"[*American Naturalist*] welcomes manuscripts that develop new conceptual syntheses, especially those combining verbal or mathematical theory with new empirical information of general significance."
Ecology	*Articles:* "Articles describing significant original research comprise the core of the journal." *Reports:* "Reports are expected to disclose new and exciting work in a concise format. These papers should present results that substantially advance a field or overturn existing ideas." *Notes:* "Present significant new observations and methodological advances."
Evolution	"[W]ell-written papers that represent significant new findings, are of general interest, and are placed in a general context are most likely to be published in *Evolution*."
J. of Animal Ecology	"[P]ublishes the best in original research on any aspect of animal ecology."
Nature	"The initial criteria for a paper to be sent for peer-review are that the results seem novel, arresting (unexpected or surprising), and that the work seems broadly significant outside the field."
Science	"Is your paper one of the best you have ever produced? Will it have a big impact in your field? Will scientists in related fields be interested in the results? Will it surprise the reader? Does it overturn conventional wisdom?"

TABLE 2b Portions of instructions to referees from various journals

Journal	Instructions to Referees
American Naturalist	Will this paper (score on a scale from 1−5): "c. Pose a new and significant problem or introduce a novel subject to the readership. d. Change the way people think about the topic of the manuscript. e. Confirm or refute an unverified theoretical principle or a previously unsupported/weakly supported generalization."
Ecology	Rating of manuscript (numerical score): "Importance to our readers. Originality of the research."
Evolution	"Does the paper contain new data or new ideas? (yes/no)"
Oecologia	"First, is the science as such sound ... ? Second ... how do you rate its interest value to *Oecologia*? We ... try to select papers which are not merely confirmatory, but make some novel contribution ..."
Oikos	(choose one) "1. Excellent, opens a new and significant area of research; 2. Very good. Makes a conceptual advance in an established field; 3. Good. Adds significantly to knowledge in an established field; 4. Could potentially reach standard 1 or 2 or 3 (indicate which) after revisions; 5. Sound but unexciting routine work that makes no significant contribution to knowledge and no conceptual advance."
Nature	"If the conclusions are not original, it would be very helpful if you could provide relevant references."
Science	"Priority is given to papers that reveal novel concepts of interdisciplinary interest". "In selecting papers for publication, the editors give preference to those papers with novelty and general significance."

DNA sequence (51)—had wide enough appeal. Is the implication here that few data in ecology or evolution are so critical as to require verification?

How To Ameliorate the Impact of Quasireplication and Selective Reporting?

What can be changed to reduce the effects of selective reporting and enhance the stature of formal replicative research? Five changes—in increasing order of difficulty to implement—would help.

(1) Journal Format One minor change to journal formats would have immediate results and be easy to implement. A special category called *Replications*, in which only fully independent replications would be published, offers four advantages. First, it would give greater prominence to truly replicated studies and would presumably encourage more biologists to conduct them. Second, it would ensure that authors, referees, and editors all gave due recognition to the value of formal replications. Third, it would place replicative studies in a defined area of the journal so that readers interested only in novel or original research would not

be distracted. Finally, if authors knew that others might attempt to replicate particularly striking claims because replicative studies were actually encouraged by journals, they might be more cautious about rushing flashy but preliminary results through to publication.

(2) Editorial Policy Regarding Replication Journal editorial policies could readily be revised to recognize the value of replicate studies without them becoming a burden to readers. Authors of replicate studies would be obliged to make the case that the study being replicated was central to a developing generalization or dogma, in the same way they are now obliged to make the case that their research is original. Purely pedestrian replication of peripheral studies should garner no more respect than purely pedestrian quasireplications.

(3) Editorial Policy Regarding Statistical Significance Journal editorial policies could include two additional yardsticks for judging the robustness of a result: (*a*) a quantitative sliding scale of statistical significance that depends on sample size, and (*b*) a qualitative nonsignificance test.

The problem of selective reporting is a simple one: As sample size decreases, an author's decision to publish will be more and more influenced by the statistical significance of the result. To discourage the publication of exploratory results that happen to reach statistical significance due to chance, the α level for a result to be considered statistically significant should be more extreme for small than for large samples. For example, for a correlation coefficient, set $\alpha = 0.001$ for sample sizes less than 20, set $\alpha = 0.01$ for sample sizes of 20–100, and retain the usual convention of $\alpha = 0.05$ for sample sizes exceeding 100. Such a rule would go a long way toward reducing the publication of spuriously significant results based on small sample size. What is considered a small sample size would depend on the type of statistical test, but standard meta-analytic techniques (91) allow other statistics to be converted to a common "effect size" (e.g., a correlation coefficient; 7) for easier judgment.

In addition, referees and editors could apply a rough rule of thumb: the nonsignificance test. The nonsignificance test would serve as a kind of litmus test of the strength of a significant result: Is the hypothesis, and sampling or experimental design, sufficiently robust that the results would seem worthy of publication if not significant statistically? In other words, if the primary result was clearly nonsignificant at $P = 0.5$, would the study still seem worthy of publication? If the answer to the nonsignificance test was "yes," then a weakly significant result would be worth reporting. However, if the answer was "no," then a weakly significant result would seem dubious at best.

Both of these suggestions are related to power analysis, but retrospective power analysis suffers from a number of problems (103) and seems unlikely to offer a simple solution. A power analysis asks, If the parametric value of a statistical descriptor for a particular population is truly nonzero, how often would it be found to be significantly different from zero for a given sampling error and sample size?

Unfortunately, retrospective power analyses—those asking about the power of a final result—appear to offer no solution. First, for many statistics a result that is just barely significant (e.g., $P = 0.049$) will always yield a power of approximately 0.5 because 50% of replicate samples from the same population would be higher than this value (and thus significant statistically), and 50% would be lower (and thus nonsignificant), no matter what the sample size. Because it is simply inversely related to the original P value, retrospective estimates of power offer no additional information (103). Second, the confidence intervals on estimates of power can be large (103), thus rendering them uninformative.

(4) Undergraduate and Graduate Training Graduate program coordinators and supervisors should consider encouraging graduate students, as part of a graduate degree program, to conduct at least one formal replicative study. Clearly these would have to be combined with original research so that students could also demonstrate their creativity and ability to tackle novel problems. Here again, a *Replications* section in premier journals would help legitimize and encourage replicative research, and allow students to be recognized for a well-replicated study as much as for a wholly novel one.

(5) Research Funding Priorities Research funding agencies could create a special funding category called *Critical Replications*. Not only would this reward replicative research directly, it would also allow a registry of formally replicated studies to be maintained. As in medical research (100), such registries ensure that reviewers and meta-analysts could directly assess the magnitude of selective reporting. In addition, such funding could be awarded with the condition that it must be published before the same investigators would be eligible to apply for any subsequent funding from the *Critical Replications* fund.

Is "True" Replication Possible?

Some will object that true replication of ecological or evolutionary studies is not possible even in principle. Even where a study is replicated with the same population of the same species using the same protocol, many other factors may not be controllable (weather, genetic makeup of a population, population density of the study organism or of other organisms that might affect the outcome of the study, other historical effects, etc). These factors might also affect replications attempted with a different population of the same species and therefore make interpretation troublesome. This inability to truly replicate a previous study is, undoubtedly, a significant problem.

But is this not precisely why replication is valuable in the first place—to judge just how repeatable (and therefore presumably biologically significant) a result is? Should we not care about how large the effect of uncontrolled variation is on the magnitude or clarity of a particular cherished result? Although it may take different forms, uncontrollable variation is a fact of life for all scientific research.

Surely what matters is either (*a*) how repeatable a result is in the face of uncontrolled variation, or (*b*) how sensitive a result is to specific uncontrollable variables. Clearly, "there is a vast amount of extra information available from repeated experimentation (generality of circumstances; variation in intensity; consistency over seasons, etc.)" (106a). The ineluctable variability of natural systems is a very real and sometimes fascinating aspect of the biology. We ignore it at our peril.

Selective Reporting Among Replicated Studies

Will increased replication eliminate the problem of selective reporting? Of course not. Those with a strong desire to confirm an earlier result will be more inclined to report a positive than a negative outcome. Similarly, those with a vested interested in contradicting a previous claim will be more inclined to publish contrary results. Such biases, even among replicated studies, may be widespread in particularly litigious areas [e.g., see the debate over the effects of second-hand smoke between Givens et al (43) and commentators (14)].

Nonetheless, replicative studies, at the very least, provide some estimate of the among-study variance—due both to genuine sampling error and to selective reporting—that is inevitably present in the scientific enterprise. Quasireplication will never allow the among-study variance to be separated from the among-taxon or among-system variances. Furthermore, if effects are so weak as to yield contradictory results in the hands of different investigators, perhaps it is time to acknowledge that the phenomenon under study may be of little biological significance. For example, the inability of multiple labs to obtain the same results with the same protocol when measuring the anxiety levels of the same strains of laboratory mice (34) should clearly give serious pause to those who wish to study the genes responsible for anxiety.

CONCLUSION

Few would dispute the enviable success that some disciplines in molecular biology have achieved—a success anticipated over 25 years ago (85). Molecular biologists' ability to weed out flawed methods or results via replication has undoubtedly promoted this sustained success. Clearly, ecologists and evolutionary biologists must face an ugly fact: Such success will elude our grasp until formal replication of others' work is embraced as a routine and respected element of research. As Bacon (9) observed, "truth will sooner come out from error than from confusion."

Quasireplication alone will not suffice. It is so vulnerable to selective reporting that it will as likely reinforce trendy notions as it will strengthen genuine biological generalizations. Without true replication we will never know which cherished generalizations are valid and which are the unfortunate consequence of collective wishful thinking re-enforced by an injudicious faith in statistics. Pity, eh?

ACKNOWLEDGMENTS

Many individuals aided me with this review and I am grateful to all for their efforts. Ary Hoffmann, Tim Mousseau, and Derek Roff provided digital or hard copies of their extensive tabulations of heritability estimates though, in the end, space constraints prevented me from incorporating them. I am particularly grateful to Brian Leung for sending me his detailed compilation of effect sizes for pre-1996 FA-stress and FA-fitness relations (63). Stefan Van Dongen and Robert Poulin kindly provided preprints of papers in press. Stefan Van Dongen also generously conducted the simulations that yielded Figure 9. Curt Strobeck, as always, provided a valuable sounding board for statistical issues. Jayson Gillespie carefully entered extensive tables of published results, unearthed sample sizes from original studies, and retrieved many useful papers for me. Many editors or managing editors quickly provided copies of their instructions to referees. Reuben Kaufman, Pete Palmer, Daphne Fautin, and particularly Lois Hammond and Locke Rowe, all provided helpful comments on early drafts of the paper.

LITERATURE CITED

1. Alatalo RV, Kotiaho J, Mappes J, Parri S. 1998. Mate choice for offspring performance: major benefits or minor costs? *Proc. R. Soc. Lond. B* 265:2297–301
2. Alatalo RV, Mappes J, Elgar MA. 1997. Heritabilities and paradigm shifts. *Nature* 385:402–3
3. Ankney CD. 1982. Sex ratio varies with egg sequence in lesser snow geese. *Auk* 99:662–66
4. Anscombe FJ. 1973. Graphs in statistical analysis. *Am. Stat.* 27:17–21
5. Antolin MF. 1993. Genetics of biased sex ratios in subdivided populations: models, assumptions, and evidence. *Oxford Surv. Evol. Biol.* 9:239–81
6. Arnqvist G, Rowe L, Krupa JJ, Sih A. 1996. Assortative mating by size: a meta-analysis of mating patterns in water striders. *Evol. Ecol.* 10:265–84
7. Arnqvist G, Wooster D. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. *Trends Ecol. Evol.* 10:236–40
8. Bacon F. 1891. *The Advancement of Learning*. Oxford: Clarendon. 4th ed.
9. Bacon F. 1960. *The New Organon and Related Writings*. New York: Liberal Arts
10. Becker BJ. 1994. Combining significance levels. See Ref. 25, pp. 215–30
11. Becker BJ. 1998. Mega-review: books on meta-analysis. *J. Educ. Behav. Stat.* 23:77–92
12. Begg CB. 1994. Publication bias. See Ref. 25, pp. 399–409
13. Begg CB, Berlin JA. 1988. Publication bias: a problem in interpreting medical data. *J. R. Stat. Soc.* A151:419–63
14. Begg CB, DuMouchel W, Harris J, Dobson A, Dear K, et al. 1997. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate—comments and rejoinders. *Stat. Sci.* 12:241–50
15. Blank JL, Nolan V. 1983. Offspring sex ratio in red-winged blackbirds is dependent on maternal age. *Proc. Natl. Acad. Sci. USA* 80:6141–45

16. Brookfield J. 1999. Allozymes again. *Evolution* 53:632–34
17. Cappelleri JC, Ioannidis JPA, Schmid CH, deFerranti SD, Aubert M, et al. 1996. Large trials vs meta-analysis of smaller trials—How do their results compare? *J. Am. Med. Assoc.* 276:1332–38
18. Chamberlin TC. 1897. The method of multiple working hypotheses. *J. Geol.* 5:837–48
19. Cleary RJ, Casella G. 1997. An application of Gibbs sampling to estimation in meta-analysis: accounting for publication bias. *J. Educ. Behav. Stat.* 22:141–54
20. Clutton-Brock TH. 1986. Sex ratio variation in birds. *Ibis* 128:317–29
21. Cockburn A. 1990. Sex ratio variation in Marsupials. *Aust. J. Zool.* 37:467–79
22. Cooke F, Harmsen R. 1983. Does sex ratio vary with egg sequence in lesser snow geese? *Auk* 100:215–17
23. Cooper H. 1979. Statistically combining independent studies: a meta-analysis of sex differences in conformity research. *J. Pers. Soc. Psychol.* 37:131–46
24. Cooper H, DeNeve K, Charlton K. 1997. Finding the missing science: the fate of studies submitted for review by a human subjects committee. *Psychol. Meth.* 2:447–52
25. Cooper H, Hedges LV, eds. 1994. *The Handbook of Research Synthesis*. New York: Russel Sage Found.
26. Csada RD, James PC, Espie RHM. 1996. The 'file drawer problem' of nonsignificant results: Does it apply to biological research? *Oikos* 76:591–93
27. Deeks JJ. 1998. Systematic reviews of published evidence: miracles or minefields? *Ann. Oncol.* 9:703–9
28. Devlin B, Daniels M, Roeder K. 1997. The heritability of IQ. *Nature* 388:468–71
29. Dickersin K. 1997. How important is publication bias? A synthesis of available data. *AIDS Educ. Preven.* 9:15–21
30. Egger M. 1998. Under the metarescope: potential and limits of meta-analyses. *Schweizerische Medizinische Wochenschrift* 128:1893–1901
31. Egger M, Smith GD. 1998. Meta-analysis: bias in location and selection of studies. *Br. Med. J.* 316:61–66
32. Egger M, Smith GD, Schneider M, Minder C. 1997. Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J.* 315:629–34
33. Elner RW, Vadas RLS. 1990. Inference in ecology: the sea-urchin phenomenon in the northwestern Atlantic. *Am. Nat.* 136:108–25
34. Enserink M. 1999. Fickle mice highlight test problems. *Science* 284:1599–600
35. Falconer DS. 1981. *Introduction to Quantitative Genetics*. New York: Longman
36. Festa-Bianchet M. 1996. Offspring sex ratio studies of mammals: Does publication depend upon the quality of the research or the direction of the results? *Ecoscience* 3:42–44
37. Fiala KL. 1981. Sex ratio constancy in the red-winged blackbird. *Evolution* 35:898–910
38. Fields SJ, Spiers M, Herschkovitz I, Livshits G. 1995. Reliability of reliability coefficients in the estimation of asymmetry. *Am. J. Phys. Anthropol.* 96:83–87
39. Finney DJ. 1995. A statistician looks at meta-analysis. *J. Clin. Epidemiol.* 48:87–103
40. Fisher RA. 1958. *The Genetical Theory of Natural Selection*. New York: Dover. 2nd ed.
41. Frank SA. 1990. Sex allocation theory for birds and mammals. *Annu. Rev. Ecol. Syst.* 21:13–55
42. Gangestad SW, Thornhill R. 1999. Individual differences in developmental precision and fluctuating asymmetry: a model and its implications. *J. Evol. Biol.* 12:402–16
42a. Gardner M. 1990. *The New Ambidextrous Universe*. New York: Freeman. 3rd ed.
43. Givens GH, Smith DD, Tweedie RL.

1997. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Stat. Sci.* 12:221–40
44. Gowaty PA. 1993. Differential dispersal, local resource competition, and sex ratio variation in birds. *Am. Nat.* 141:263–80
45. Gurevitch J, Morrow LL, Wallace A, Walsh JS. 1992. A meta-analysis of competition in field experiments. *Am. Nat.* 140:539–72
46. Hagen DW. 1973. Inheritance of number of lateral plates and gill rakers in *Gasterosteus aculeatus*. *Heredity* 30:303–12
47. Hamilton WJ, Poulin R. 1997. The Hamilton and Zuk hypothesis revisited: a meta-analytical approach. *Behaviour* 134:299–320
48. Harmsen R, Cooke F. 1983. Binomial sex-ratio distribution in the lesser snow goose: a theoretical enigma. *Am. Nat.* 121:1–8
49. Harrington A, ed. 1999. *The Placebo Effect: An Interdisciplinary Exploration*. Cambridge: Harvard Univ. Press
50. Hedges LV. 1992. Modeling publication selection effects in meta-analysis. *Stat. Sci.* 7:246–55
51. Höss M. 2000. Neanderthal population genetics. *Nature* 404:453–54
52. Houle D. 1998. High enthusiasm and low *R*-squared. *Evolution* 52:1872–76
53. Howe HF. 1977. Sex ratio adjustment in the common grackle. *Science* 198:744–46
54. Hu S, Pattatucci AML, Patterson C, Li L, Fulker DW, et al. 1995. Linkage between sexual orientation and chromosome Xq28 in males but not in females. *Nat. Genet.* 11:248–56
55. Hunt M. 1997. *How Science Takes Stock: The Story of Meta-Analysis*. New York: Russell Sage Found.
56. Jarvinen 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* 1333:62–66
57. Jones DR. 1995. Meta-analysis: weighing the evidence. *Stat. Med.* 14:137–49
58. Komdeur J, Daan S, Tinbergen J, Mateman C. 1997. Extreme adaptive modification of sex ratio of the Seychelles warbler's eggs. *Nature* 385:522–25
59. Krackow S. 1995. Potential mechanisms for sex ratio adjustment in mammals and birds. *Biol. Rev.* 70:225–41
60. Kuhn TS. 1962. *The Structure of Scientific Revolutions*. Chicago: Univ Chicago Press. 2nd ed.
61. Lau J, Ioannidis JPA, Schmid CH. 1997. Quantitative synthesis in systematic reviews. *Ann. Intern. Med.* 127:820–26
62. Leamy L. 1993. Morphological integration of fluctuating asymmetry in the mouse mandible. *Genetica* 89:139–53
63. Leung B, Forbes MR. 1996. Fluctuating asymmetry in relation to stress and fitness: effects of trait type as revealed by meta-analysis. *Ecoscience* 3:400–13
64. Light RJ, Pillemer DB. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge: Harvard Univ. Press
65. Ludwig W. 1932. *Das Rechts-Links Problem im Teirreich und beim Menschen*. Berlin: Springer
66. Lynch M, Walsh B. 1997. *Genetics Analysis of Quantitative Traits*. Sunderland, MA: Sinauer
67. Magnusson WE. 2000. Error bars: Are they the king's clothes? *Bull. Ecol. Soc. Am.* 147–50
68. Mann C. 1990. Meta-analysis in the breech. *Science* 249:476–80
69. Mather K. 1953. Genetical control of stability in development. *Heredity* 7:297–336
70. Mather K, Jinks JL. 1977. *Introduction to Biometrical Genetics*. New York: Chapman & Hall
71. Møller AP. 1992. Female swallow preference for symmetrical male sexual ornaments. *Nature* 357:238–40
72. Møller AP. 1999. Asymmetry as a predictor of growth, fecundity and survival. *Ecol. Lett.* 2:149–56

73. Møller AP, Swaddle JP. 1997. *Developmental Stability and Evolution*. Oxford: Oxford Univ. Press
74. Møller AP, Thornhill R. 1998. Bilateral symmetry and sexual selection: a meta-analysis. *Am. Nat.* 151:174–92
75. Mosteller F, Colditz GA. 1996. Understanding research synthesis (meta-analysis). *Annu. Rev. Public Health* 17:1–23
76. Newton I, Marquiss M. 1979. Sex ratio among nestlings of the European sparrowhawk. *Am. Nat.* 113:309–15
77. Normand S-LT. 1999. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.* 18:321–59
78. Palmer AR. 1994. Fluctuating asymmetry analyses: a primer. In *Developmental Instability: Its Origins and Evolutionary Implications*, ed. TA Markow, pp. 335–64. Dordrecht, Netherlands: Kluwer
79. Palmer AR. 1996. Waltzing with asymmetry. *BioScience* 46:518–32
80. Palmer AR. 1999. Detecting publication bias in meta-analyses: a case study of fluctuating asymmetry and sexual selection. *Am. Nat.* 154:220–33
81. Palmer AR, Hammond LM. 2000. The Emperor's codpiece: a post-modern perspective on biological asymmetries. *Int. Soc. Behav. Ecol. Newslett.* 12: In Press
82. Palmer AR, Strobeck C. 1986. Fluctuating asymmetry: measurement, analysis, patterns. *Annu. Rev. Ecol. Syst.* 17:391–421
83. Phillips PC. 1998. Designing experiments to maximize the power of detecting correlations. *Evolution* 52:251–55
84. Pigliucci M, Kaplan J. 2000. The fall and rise of Dr. Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends Ecol. Evol.* 15:66–70
85. Platt JR. 1964. Strong inference. *Science* 146:347–53
86. Poulin R. 2000. Manipulation of host behaviour by parasites: a weakening paradigm? *Proc. R. Soc. Lond. B.* 267:787–92
87. Quinn JF, Dunham AE. 1983. On hypothesis testing in ecology and evolution. *Am. Nat.* 122:602–7
88. Rice G, Anderson C, Risch N, Ebers G. 1999. Male homosexuality: absence of linkage to microsatellite markers at Xq28. *Science* 284:665–67
89. Rohlf FJ, Sokal RR. 1995. *Statistical Tables*. San Francisco: Freeman. 3rd ed.
90. Rosenthal R. 1979. The "file drawer problem" and tolerance for null results. 86:638–41
91. Rosenthal R. 1991. *Meta-Analytic Procedures for Social Research*. Beverly Hills: Sage
92. Roughgarden J. 1983. Competition and theory in community ecology. *Am. Nat.* 122:583–601
93. Salt GW. 1983. Roles: their limits and responsibilities in ecological and evolutionary research. *Am. Nat.* 122:697–705
94. Sheldon BC. 1998. Recent studies of avian sex ratios. *Heredity* 80:397–402
95. Simberloff D. 1983. Competition theory, hypothesis testing, and other community ecological buzzwords. *Am. Nat.* 122:626–35
96. Simmons LW, Tomkins JL, Kotiaho JS, Hunt J. 1999. Fluctuating paradigm. *Proc. R. Soc. Lond. B* 266:593–95
97. Smith BH, Garn SM, Cole PE. 1982. Problems of sampling and inference in the study of fluctuating dental asymmetry. *Am. J. Phys. Anthropol.* 58:281–89
98. Sokal RR, Rohlf FJ. 1995. *Biometry*. New York: Freeman. 3rd ed.
99. Soulé ME, Baker B. 1968. Phenetics of natural populations. IV. The populations asymmetry parameter in the butterfly *Coenonympha tullia*. *Heredity* 23:611–14
100. Stern JM, Simes RJ. 1997. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *Br. Med. J.* 315:640–45
101. Strong DR. 1980. Null hypotheses in ecology. *Synthese* 43:271–85

102. Strong DR. 1983. Natural variability and the manifold mechanisms of ecological communities. *Am. Nat.* 122:636–60
103. Thomas L. 1997. Retrospective power analysis. *Conserv. Biol.* 11:276–80
104. Thornhill R, Møller AP. 1998. The relative importance of size and asymmetry in sexual selection. *Behav. Ecol.* 9:546–51
105. Thornhill R, Møller AP, Gangestad SW. 1999. The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. *Am. Nat.* 154:234–41
106. Thornhill R, Sauer P. 1992. Genetic sire effects on the fighting ability of sons and daughters and mating success of sons in a scorpionfly. *Anim. Behav.* 43:255–64
106a. Underwood AJ. 1999. Publication of so-called 'negative' results in marine ecology. *Mar. Ecol. Prog. Ser.* 191:307–9
107. Van Dongen S. 1998. How repeatable is the estimation of developmental stability by fluctuating asymmetry? *Proc. R. Soc. Lond. B* 265:1423–27
108. Van Dongen S. 2000. The heritability of fluctuating asymmetry: a Bayesian hierarchical model. *Ann. Zool. Fennici.* 37:15–23
109. Van Dongen S, Lens L. 2000. The evolutionary potential of developmental stability. *J. Evol. Biol.* 13:326–35
110. van Schaik CP, Hrdy SB. 1991. Intensity of local resource competition shapes the relationship between maternal rank and sex ratios at birth in cercopithecine primates. *Am. Nat.* 138:1555–62
111. Van Valen L. 1962. A study of fluctuating asymmetry. *Evolution* 16:125–42
112. Vickers A, Goyal N, Harland R, Rees R. 1998. Do certain countries produce only positive results? A systematic review of controlled trials. *Contr. Clin. Trials* 19:159–66
113. Vollestad LA, Hindar K, Moller AP. 1999. A meta-analysis of fluctuating asymmetry in relation to heterozygosity. *Heredity* 83:206–18
114. Wang MC, Bushman BJ. 1998. Using the normal quantile plot to explore meta-analytic data sets. *Psychol. Meth.* 3:46–54
115. Weatherhead P. 1983. Secondary sex ratio adjustment in red-winged blackbirds (*Agelaius phoeniceus*). *Behav. Ecol. Sociobiol.* 12:57–61
116. Weigensberg I, Roff DA. 1996. Natural heritabilities: Can they be reliably estimated in the laboratory? *Evolution* 50: 2149–57
117. Whitlock M. 1996. The heritability of fluctuating asymmetry and the genetic control of developmental stability. *Proc. R. Soc. Lond. B* 263:849–53
118. Whitlock M. 1998. The repeatability of fluctuating asymmetry: a revision and extension. *Proc. R. Soc. Lond. B* 265:1429–31
119. Williams GC. 1979. On the question of adaptive sex ratio in outcrossed vertebrates. *Proc. R. Soc. Lond. B* 205:567–80
120. Wilson EE. 1975. *Sociobiology*. Cambridge, MA: Harvard Univ. Press
121. Woods RE, Hercus MJ, Hoffmann AA. 1998. Estimating the heritability of fluctuating asymmetry in field *Drosophila*. *Evolution* 52:816–24
122. Woods RE, Sgro CM, Hercus MJ, Hoffmann AA. 1999. The association between fluctuating asymmetry, trait variability, trait heritability, and stress: a multiply replicated experiment on combined stresses in *Drosophila melanogaster*. *Evolution* 53:493–505