Oecologia (2002) 132:492–500
DOI 10.1007/s00442-002-0952-2

METHODS

Anders Pape Møller · Michael D. Jennions

How much variance can be explained by ecologists and evolutionary biologists?

Received: 12 June 2001 / Accepted: 18 April 2002 / Published online: 19 July 2002

Abstract The average amount of variance explained by the main factor of interest in ecological and evolutionary studies is an important quantity because it allows evaluation of the general strength of research findings. It also has important implications for the planning of studies. Theoretically we should be able to explain 100% of the variance in data, but randomness and noise may reduce this amount considerably in biological studies. We performed a meta-analysis using data from 43 published meta-analyses in ecology and evolution with 93 estimates of mean effect size using Pearson's *r* and 136 estimates using Hedges' *d* or *g*. This revealed that (depending on the exact analysis) the mean amount of variance (r^2) explained was 2.51–5.42%. The various 95% confidence intervals fell between 1.99 and 7.05%. There was a strongly positive relationship between the fail-safe number (the number of null results needed to nullify an effect) and the coefficient of determination (r^2) or effect size. Analysis at the level of individual tests of null hypotheses showed that the amount of variance key factors explained differed among fields with the largest amount in physiological ecology, lower amounts in ecology and the lowest in evolutionary studies. In all fields though, the hypothesized relationship (e.g. main effect of a fixed treatment) explained little of the variation in the trait of interest. Our finding has important implications for the interpretation of scientific studies. Across studies, the average effect size reported is between Pearson *r*=0.180 and 0.193 and Hedges' *d*=0.631 and 0.721. Thus the average sample sizes needed to conclude that a particular relationship is absent with a power of 80% and α=0.05 (two-tailed) are considerably larger than usually recorded in studies of evolution and ecology. For example, to detect *r*=0.193, the required sample size is 207.

Keywords Ecology · Effect size · Evolution · Meta-analysis · Sample size

A.P. Møller (✉)
Laboratoire d'Ecologie Evolutive Parasitaire, CNRS UMR 7103, Université Pierre et Marie Curie, 7 quai St. Bernard, Case 237, 75252 Paris Cedex 05, France
e-mail: amoller@snv.jussieu.fr
Tel.: +33-1-44272594, Fax: +33-1-44273516

M.D. Jennions
Smithsonian Tropical Research Institute, Unit 0948, APO AA 34002–0948, USA

M.D. Jennions
School of Botany and Zoology, Australian National University, Canberra, A.C.T. 0200, Australia

Introduction

Effect size is a standardized measure of the magnitude of a relationship. Standard measures of effect size include Pearson's product-moment correlation coefficient *r* and change measured in units of standard deviations (e.g. Hedges' *d* or *g*) (Hedges and Olkin 1985; Rosenthal 1991, 1994; Cooper and Hedges 1994). Here we use Pearson's *r* and Hedges' *d* as measures of effect size. Pearson's *r* has the appealing, intuitive feature that its squared value represents the amount of variance explained by the predictor variable (Rosenthal 1991, 1994). As a simple rule of thumb, Cohen (1988) suggested that a 'small' effect has a mean correlation coefficient of 0.10 (i.e. explains 1% of the variance since r^2=1%), a 'medium' effect has a coefficient of 0.30 (i. e. explains 9% of the variance), and a 'large' effect has a coefficient of 0.50 (i.e. explains 25% of the variance). Similarly, a Hedges' *d* of 0.2 is considered a small effect, *d*=0.5 an intermediate effect and *d*=0.8 a large effect (Cohen 1988). Biologists performing power analyses tend to present results assuming unknown effects are of 'medium' strength. This is an arbitrary categorization, but still useful because it allows assessment of the magnitude of research findings. When comparing research, we are almost always talking about magnitudes rather than absolute estimates.

When first conducting research many graduate students are disappointed when they encounter the fact that biologists explain so little of the variance in their data.

That is particularly true if we investigate the generality of a research finding across a large number of studies. Thus, the naive question is as follows: Can we ever explain 100% of the variance? The obvious answer is no, and there are several reasons why that is the case. In particular, biology differs from many other subjects in the natural sciences by being considerably more complex, with consequences for the amount of variation that can be explained by observational or experimental studies. First, biological systems are not "perfect" because adaptation is not "perfect". There is always a certain lag because of phylogenetic constraint or selection pressures preceding responses to selection. Second, "randomness" caused by unpredictable physical properties of the environment may considerably reduce the amount of variance explained. Third, organisms usually balance their responses in relation to many different factors (e.g. size, temperature, condition, predation risk, age), and biologists are rarely able to measure more than a few of these. So "noise" caused by the effects of "confounding" variables will tend to render relationships "blurry" when we only examine the relationship between two traits and fail to control adequately for these other factors. Fourth, measurement accuracy affects the amount of variance explained because a coarser yardstick provides a rougher measure of "reality". Accurate measurements of many biological traits are extremely difficult (e.g. size-based mortality, behavioral propensities) not least because they usually vary geographically and temporally. Fifth, all organisms by definition have an evolutionary past that affects their ability to adapt to current conditions. Hence, current levels of adaptation will be affected by evolutionary history. Sixth, the activities of one individual may affect those of others, so neither may obtain a solution that can be considered optimal; there are limits to optimality (Maynard Smith 1978).

But is it really important to know how much variance can generally be explained by key factors of interest to ecologists and evolutionary biologists? We think so. Knowing the amount of variance usually explained by the fixed factors, on which we focus, puts studies into sharper perspective. For example, a study showing that a treatment or covariate only explains 5% of the variance in a character of interest may superficially indicate a weak relationship if we use 100% as our yardstick. However, if the main factors biologists examine rarely explain more than 10% of the variance in ecological and evolutionary studies, these seemingly inconsequential factors suddenly become far more important. Likewise, if effects generally are only a small amount of variance explained, this raises the importance of using powerful research synthesis methods such as meta-analysis.

Of course, in an evolutionary context, a difference in phenotype of 0.1 standard deviations or less per generation can be extremely important. These small differences readily "change a mouse into an elephant". Studies of the fossil record typically measure a change in phenotype per generation of far smaller magnitude (Simpson 1944, 1949; Haldane 1949; Ridley 1993). Higher rates are more common on micro-evolutionary time scales although rates of evolutionary change are still generally very small (Hendry and Kinnison 1999). For this reason, we should not immediately dismiss small effects as inconsequential. A distinction should be drawn between short term predictability and long-term effects.

Some factors that reduce the amount of variance explained can be controlled for experimentally. Experiments identify causal relationships by directly manipulating one or more factors, while holding others constant. Furthermore, experiments usually include relatively more extreme phenotypes or situations than found naturally. This allows for a clearer assessment of potential relationships. The effectiveness of an experimental approach, measured as the increase in the amount of variance explained, has, however, not been generally examined.

For studies based on small sample sizes individual estimates of effect sizes show a larger range of values than they do when sample sizes are large. This results in a so-called "funnel" graph that converges towards the 'true' effect size as sample size increases and variance in effect size estimates decreases (Light and Pillemer 1984; Palmer 1999; Møller and Jennions 2001). This is the main reason for weighting estimates of effect size by the sample size on which they are based (or, more technically-speaking, by the inverse of the variance in the estimate) when calculating mean effect size in a meta-analysis.

Publication bias has traditionally been assessed by examining the robustness of general research findings. If a large number of additional studies with an average effect of zero are needed to nullify the mean effect size of a meta-analysis, it is safe to conclude that the generalization is robust (Rosenthal 1991, p 104). The so-called "fail-safe number" is the number of such null results. Alternatively, it can be considered to represent the number of unpublished results resting in the file drawers of scientists. If large, we can conclude that it is unlikely that publication bias will alter our main findings (unless results opposite in direction to the reported mean effect are less likely to be published). Rosenthal (1991) suggested that a fail-safe number 5 times larger than the sample size plus 10 indicates a robust result. Most recently Gurevitch and Hedges (1999) have suggested that reliance on the fail-safe number is an appropriate step to resolve potential problems of publication bias.

The natural sciences are often described as the exact sciences. This is certainly the case for mathematics, physics and chemistry, but much less so for most fields in biology. Biologists deals with living organisms under the influence of numerous biotic and abiotic interactions that alone, and in combination, influence whichever factor is under investigation. Thus the predictive power of most fields of biology is considerably weaker than that in the more exact natural sciences. Although this is a common notion prevalent in general books about science and the philosophy of science, there is, to the best of our knowledge, no study investigating the general level of predictability in biology as determined from a review of meta-analyses in different fields.

In this study we present a meta-analysis of meta-analyses in ecology and evolution to assess the amount of variance explained by the main factors that researchers have focused on. Although meta-analyses have limitations, at least they generally explicitly state the criteria used to include studies, while this is rarely the case for narrative reviews. First, we quantify the general strength of relationships studied by biologists. Second, we contrast the amount of variance explained in experiments as opposed to observational studies by using estimates of the mean effect size from the two categories of studies taken from the same original meta-analysis. This allows quantification of the effectiveness of experiments, and an estimate of the amount of variance explained by the removal of confounding variables and increased phenotypic variation. Third, we quantify the relationship between the amount of variance explained by a factor and the associated fail-safe number to test the prediction that a larger mean amount of variance explained gives rise to a larger fail-safe number. Finally, we quantify differences in effect size between different areas of biology, namely physiology, ecology and evolution. We expect such fields to differ in the amount of variance explained because the complexity of external factors affecting a given relationship increases from studies of internal physiological processes over ecological studies to evolutionary studies investigating many different species. To conclude we use our estimates of mean effect sizes to determine the sample sizes needed to reach the conclusion that a particular study does not show an average effect with a power of 80%. Interestingly, these sample sizes turn out to be considerably larger than those generally recorded in studies of ecology and evolution.

Materials and methods

We made an extensive survey of the ecological and evolutionary literature for meta-analyses that could be used to estimate mean effect sizes up until the end of 2000. We examined the journals *American Naturalist*, *Animal Behaviour*, *Behavioral Ecology*, *Behavioral Ecology and Sociobiology*, *Ecological Monographs*, *Ecology*, *Evolution*, *Evolutionary Biology*, *Journal of Evolutionary Biology*, and *Quarterly Review of Biology*. We also entered the phrase 'meta-analy*' into the electronic database *WebSpiris* to search for papers where this term occurred in the title or abstract. We then examined the titles of all the papers listed and directly inspected any that seemed to fall into the field of evolutionary and ecological biology (most 'hits' were from the medical or social sciences). Furthermore, we contacted a number of colleagues who had used meta-analyses in their research to locate 'in press' studies. We identified a total of 43 meta-analyses that form the basis of the present analysis. The two effect sizes we used are Pearson's correlation coefficient *r* and Hedges' *d*. A few studies reported statistics that could not readily be transformed into either effect size (e.g. response ratios); or did not present sufficient raw data for us to recalculate Pearson's *r* or Hedges' *d* (usually because variances were not given); or they used meta-analysis as a tool in more complex studies (Warwick and Clarke 1993; Davis 1993; Osenberg et al. 1997). We explicitly excluded meta-analysis studies looking at genetic heritabilities because there is a clear publication bias; negative heritabilities are almost never reported (Palmer 2000). It is also unclear whether h^2 itself or an effect size based on the strength (rather than slope) of the relationship between relatives is a more appropriate effect size. The meta-analyses used here are listed in the Appendix. An electronic version of the entire data set is available from the first author upon request.

The choice of null hypothesis is important in any scientific inquiry (Anderson et al. 2000), and this has important implications for the subsequent statistical tests. In the meta-analyses used in the present study, the explicit null hypothesis was always that the mean effect size equaled zero. We calculated mean effect sizes in several ways:

1. The meta-analysis directly provided weighted mean r or Hedges' d (e.g. Arnqvist et al. 1996).
2. The meta-analysis provided r or Hedges' d for individual cases. We then calculated the weighted mean effect for these cases (e.g. Harper 1999).
3. Data was presented for each case in the form of means and standard deviations for two groups. We then calculated Hedges' d (e.g. Arnqvist and Nilsson 2000).
4. The meta-analysis provided weighted mean effects as Hedges' d in graphs (e.g. Curtis 1996).

Whenever we undertook further analyses, we first transformed r by means of Fisher's transformation to z-values. All mean effect sizes were given positive values for further analyses because the direction of the effect is often arbitrary (e.g. male vs female).

Each published meta-analysis usually provided several mean effect sizes because the data was divided into sub-groups. To minimize pseudo-replication we only used data from mutually exclusive groupings for each response variable of interest (e.g. birds vs insects). If the initial authors divided the data in several ways (e.g. birds vs insects and temperate vs tropical), we used the dividing factor for which the heterogeneity in effect size (Q_b) between groups was greatest. We only included mean effect sizes based on four or more studies because the asymptotic variance in the effect size z-transformed *r* is $1/(n-3)$. Another possible source of lack of independence between estimates of mean effect sizes is that some meta-analyses examined several response variables. Biologists are obviously interested in each of these response variables. It is therefore reasonable to present separate data for more than one response variable per meta-analysis. We refer to these analyses as being at the 'response variable level'. To be conservative, however, we also calculate the weighted mean effect size per published meta-analysis (i.e. one data point each). In subsequent analyses at the 'meta-analysis level' we calculated the mean effect size across the 43 meta-analyses.

For the effect size Pearson's *r*, we weighted each meta-analysis by either the average number of studies per estimate of effect size, or the total number of studies. Meta-analyses were run in Metawin 2.0 (Rosenberg et al. 2000) using mixed-models. To be conservative we present bias-corrected 95% confidence intervals calculated using bootstrapping from 999 replications. These do not require that effect sizes are parametrically distributed. For *r* we also calculated mean effect sizes and 95% confidence intervals using standard statistics. This approach treats each case or meta-analysis as being equally accurate. It tests the robustness of our results by dealing with the criticism that a few cases with very large sample sizes can generate an atypical mean estimate when performing a weighted meta-analysis. For Hedges' *d* we could not perform weighted meta-analyses because asymptotic variance in the estimate was generally unavailable. We therefore simply calculated the mean effect size at the response variable or meta-analysis level.

We classified the 43 meta-analyses as physiological, evolutionary or ecological. Studies with a mainly physiological content were put in the first category. Those that dealt with species interactions and communities were considered ecological. Evolutionary studies included studies of selection and functional aspects of behavior or behavioral ecology. We also noted whether or not the meta-analysis dealt with fluctuating asymmetry because it has been stated that the effect size in studies of asymmetry is much smaller than in other studies in evolution (Houle 1998). The classification of all meta-analyses is given in the Appendix.

Meta–analyses are problematic if null results stay unpublished (Hunter and Schmidt 1990). There are a number of methods available for testing for such bias (Light and Pillemer 1984; Vandenbroucke 1988; Berlin et al. 1989; Hedges 1992; Dear and Begg 1992; Thompson 1993; Mengersen et al. 1995; Møller and Jennions 2001). Most require relatively large data sets. One simple prediction is that, in the absence of bias, a plot of effect size against log-transformed sample size will have a funnel-shape centered around the "true" effect size. The reason for this is that variance in effect size due to sampling error decreases with increased sampling effort. Thus the reported effect sizes should be normally distributed around the mean effect with no trend in relation to sample size (Light and Pillemer 1984; Vandenbroucke 1988). In our analyses, however, we examine the absolute values for mean effect sizes because the sign of such mean effects does not make sense. Thus, in the absence of publication bias, the graph in question should resemble a half-funnel. Therefore mean effect size should increase as sample size decreases. We tested for this relationship using Spearman's correlation, after we standardized the effect size to fulfill the criteria for a rank correlation test (see Begg and Mazumdar 1994). Variance in effect size should decrease with increasing sample size. We tested this in a ratio of variance test comparing variance from studies with a sample size greater than the median than those less than the median.

Obviously, we can never know how many unpublished studies exist, but this problem can be addressed by calculating the fail-safe number of publications (Rosenthal 1991). We recorded the Rosenthal number when this was provided in the original meta-analyses. In cases where it was not provided and we could enter the original data, we calculated the fail-safe number in Metawin 2.0 using Fisher's *z* as the effect size. This only led to slight differences compared to the number obtained using the original effect size type (personal observation).

Power to detect an effect when α=0.05 (two-tailed) was calculated following Cohen (1988).

Results

Meta-analyses using Pearson's *r* as the effect size

Response variable level of analysis

Mean r^2 ranged from 0 to 48.7% across the 93 estimates of effect size *r* (Fig. 1a). More than 80% of values were smaller than 10% (*n*=76/93). While the mean value of r^2 was 5.42% (95% CI: 3.79–7.05%), the median was far lower at 2.22%. The mean value of |*r*| was 0.184 (95% CI: 0.154–0.213), thus explaining 3.39% of the variance (95% CI: 2.37–4.54%). Using meta-analysis to weight estimates of |*r*| by their variance yielded very similar results. The mean value of |*r*| was 0.180 (95% CI: 0.150–0.208) corresponding to explaining 3.24% of the variance (95% CI: 2.25–4.33%).

The median number of studies per estimate of effect size was 17. There was greater variance in mean effect sizes based on 17 or fewer studies than those based on more than 17 studies, but the difference was not significant (Variance Ratio test, *F*=1.33, *df*=45,46, *P*=0.169). There was no significant negative relationship between standardized effect size and the number of studies used to generate the effect size estimate (Fig. 2a; Spearman's *r*=–0.077, *P*=0.462, *n* =93; Power:<83%).

The fail-safe number increased with r^2, as expected if a larger effect leads to a more robust general finding. This relationship was statistically highly significant (Fig. 3; linear regression based on a log-log transformation: *F*=39.09, *df*=1,80, *P*<0.0001). The slope of this regression was 1.610 (SE 0.258), which is significantly greater than unity (*t*=2.36, *df*=80, *P*=0.021), which implies a greater than expected increase with effect size. Fail-safe number also increased with sample size (*r*=0.651, *P*<0.0001, *n*=82). The relationship between r^2 and fail-safe number remained even when sample size was included in a multiple regression (*t*=8.795, *df*=78, *P*<0.0001). It is also important to note from Fig. 3 the considerable variance in fail-safe number for a given effect size, and that small fail-safe numbers (below 100) occur even for effect sizes explaining more than 10% of the variance.

The absolute value of mean effect size did not differ among the eight meta-analyses with reported estimates for both observational and experimental studies (paired t-test, *t*=0.903, *df*=7, *P*=0.397; Power:25%). The same was true when mean effect sizes were first weighted by sampling effort (mean Cohen's *q*=0.050: 95% CI: –0.054–0.222). There was, however, a positive correla-

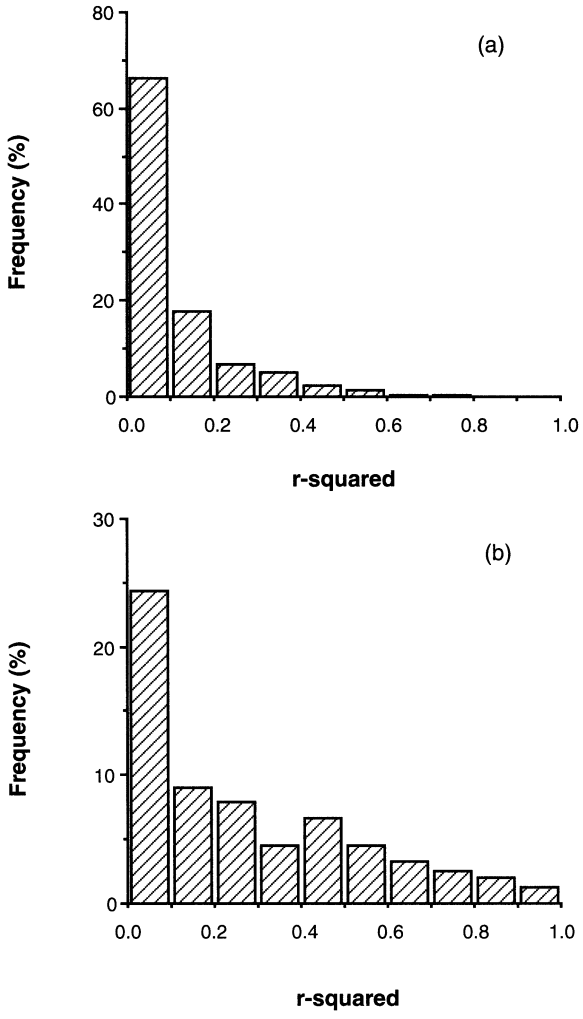

Fig. 1 Frequency distribution of **a** r^2 from meta-analyses in biology for studies with the effect size Pearson's *r* (*n*=93) and **b** Hedges' *d* (*n*=136)

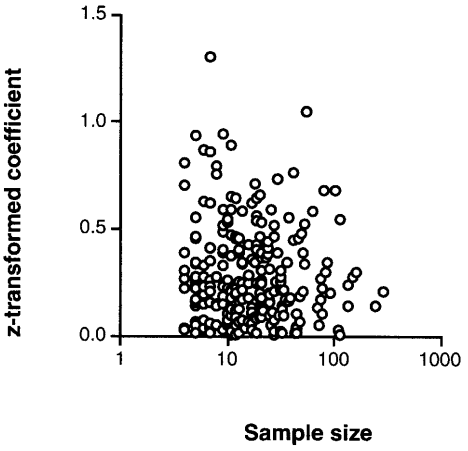

Fig. 2 The relationship between the number of studies used to calculate an effect and the absolute effect size calculated as either **a** Pearson's *r* (*n*=93) or **b** Hedges' *d* (*n*=136)

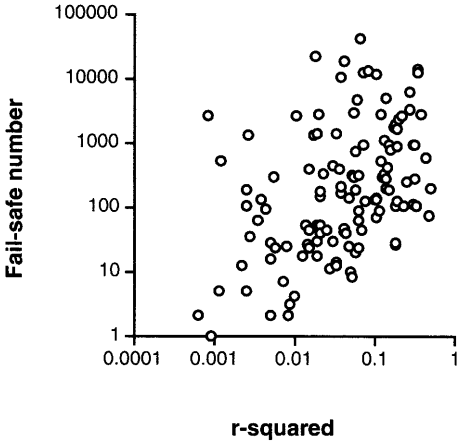

Fig. 3 The relationship between Rosenthal's Fail-safe Number and (**a**) r^2 in studies using the effect size Pearson's *r* (*n*=93) or (**b**) the effect size Hedges' *d* (*n*=136). Both relationships are positive and statistically significant (see text)

tion between observational and experimental effect size (*r*=0.746, *P*=0.034).

The mean effect size |*r*| for studies of fluctuating asymmetry did not differ from that of other evolutionary studies (ANOVA: *F*=3.018, *df*=1, 88, *P*=0.09; Power:66%). The same is true when the data is analyzed using meta-analysis with weighting for sampling effort (Q_b=0.145, *P*=0.618). For asymmetry studies |*r*|=0.198 (95% CI: 0.155–0.248) (*n*=20) for other studies |*r*|=0.183 (95% CI: 0.142–0.219) (*n*=70). Most samples (90/93) were from evolutionary studies. It was therefore impossible to compare effect size between the three main research fields.

Meta-analysis level of analysis

Mean r^2 ranged from 0.3% to 28.8% across the 22 meta-analyses with Pearson *r* as a measure of effect size. More than 90% of meta-analyses had a mean r^2 of less then 10% (*n*=21/22). The mean value of mean r^2 was 5.24% (SE=1.26); the median was lower at 4.07%. The mean value of |*r*| was 0.205 (95% CI: 0.158–0.251), thereby explaining 4.20% of the variance (95% CI: 2.50–6.30%). Using meta-analysis to weight estimates of |*r*| by their variance yielded very similar results. Weighting using mean study size per meta-analysis, the mean value of |*r*| was 0.193 (95% CI: 0.149–0.224) corresponding to explaining 2.51% of the variance (95% CI: 2.22–5.02%). Weighting using total sample size per meta-analysis, the mean absolute value of |*r*| was 0.182 (95% CI: 0.141–0.216) corresponding to explaining 3.31% of the variance (95% CI: 1.99–4.67%).

The median mean number of samples per study per meta-analyses was 22.3. There was greater variance in mean effect sizes based on 22.3 or fewer samples/study/meta-analysis than those based on more than 22.3, but not significantly so (Variance Ratio test, *F*=2.5, *df*=8,12, *P*=0.074). There was also no significant relationship between standardized effect size and the number of studies used to generate the effect size estimate (Spearman's *r*=-0.091, *P*=0.687, *n*=22; Power:<17%).

The mean effect size for studies of fluctuating asymmetry did not differ from that of other evolutionary studies (*F*=0.001, *df*=1, 19, *P*=0.98; Power:19%). The same is true when the data is analyzed using meta-analysis and weighting for sampling effort using the mean sample size per study per meta-analysis (Q_b=0.001, *P*=0.996). For asymmetry studies, |*r*| was 0.203 (95% CI: 0.178–0.252, *n*=7) and for other evolutionary studies it was also 0.203 (95% CI: 0.134–0.258, *n*=14).

Meta-analyses using Hedges' *g* or *d* as the effect size

A total of 124 of 136 samples and 17 of the 21 original meta-analyses used Hedges' *d* rather than *g* as the effect size. Hedges' *d* is simply Hedges' *g* multiplied by *J*, where $J=1-3/[4(n_c+n_E-2)-1]$. This corrects for small sample bias so Hedges' *d* is always slightly smaller than Hedges' *g*. Here we combine the two effect sizes (henceforth referred to simply as *d*). The mean sample size was greater than 20, so the difference between the two effect sizes is insignificant for the general purposes of our overview (i.e. at *n*=20, *J*=0.98). The actual means would therefore be marginally smaller if effect size had always been calculated as Hedges' *d*.

Response variable level of analysis

The value of |*d*| ranged from 0.005 to 3.416 (Fig. 1b). The mean was 0.721 (95% CI: 0.622–0.820) and the median 0.595 (*n*=136). The median number of studies per estimate of |*d*| was 13.5. There was greater variance in mean effect sizes based on 13.5 or fewer studies than those based on more than 13.5 studies (Variance Ratio test, *F*=1.47, *df*=67, 67, *P*=0.059). There was no signifi-

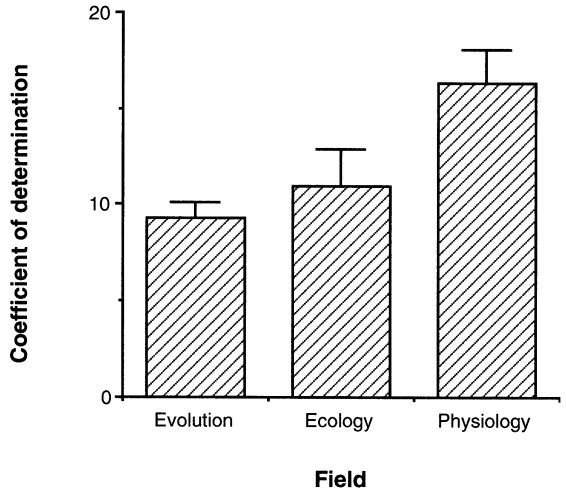

Fig. 4 Mean values of |*d*| for meta-analyses in evolution, ecology, and physiology. Values are means (SE). Sample sizes are 34, 30 and 72, respectively

cant negative relationship between |*d*| and the number of studies used to generate the effect size estimate (Fig. 2b; Spearman's *r*=0.014, *P*=0.874, *n*=136; Power:<94%).

The fail-safe number increased with |*d*|, as expected if a larger effect leads to a more robust general finding. This relationship was statistically highly significant (Fig. 3b; linear regression based on a log-log transformation: *F*=49.65, *df*=1, 52, *P*<0.0001). The slope of this regression was 5.97 (SE 0.848), which is significantly greater than unity (*t*=5.86, *df*=52, *P*<0.0001). Fail-safe number also increases with sample size (*r*=0.808, *P*<0.0001, *n*=54). The relationship between |*d*| and fail-safe number remains even when sample size is included in a multiple regression (*t*=7.36, *df*=50, *P*<0.0001).

The mean effect size for ecological, evolutionary and physiological studies differed (Fig. 4; ANOVA: *F*=3.338, *df*=2, 133, *P*=0.038). For evolutionary samples |*d*| was 0.572 (95% CI: 0.409–0.735, *n*=34), for ecological |*d*| was 0.603 (95% CI: 0.442–0.763, *n*=30) and for physiological |*d*| was 0.840 (95% CI: 0.684–0.996, *n*=72).

Meta-analysis level of analysis

The value of mean |*d*| ranged from 0.22 to 1.70. The mean was 0.631 (95% CI: 0.483–0.779) and the median was 0.577 (*n*=21). The median number of studies per estimate of |*d*| was 15.2. There was greater variance in mean effect sizes based on 15.2 or fewer studies than those based on more than 15.2 studies (Variance Ratio test: *F*=4.045, *df*=9,10, *P*=0.02). There was no significant negative relationship between |*d*| and the number of studies used to generate the effect size estimate (Spearman's *r*=-0.057, *P*=0.807, *n*=21; Power:27%).

The mean effect size for ecological, evolutionary and physiological studies did not differ (ANOVA: *F*=1.897, *df*=2, 18, *P*=0.179). Statistical power to detect a medium effect was, however, only 14%.

Discussion

What is the mean amount of variance explained by causal factors of interest to biologists? In our analyses, the weighted mean Pearson |*r*| across all estimates at the meta-analysis level was 0.19, equaling a mean coefficient of determination of 2.5%. The 95% confidence interval around this estimate fell between 2.3 and 4.3%. Looking at all the different possible analyses, the 95% confidence intervals for mean |*r*| always fell between 0.14 and 0.22 across a range of fields in biology. While other factors considered in specific studies may have explained additional variation in response variables (e.g. covariates, random effects or other fixed factors), the key factor under examination in each published meta-analysis (e.g. the effect of, say, size, symmetry or treatment) explained a relatively minor amount of the remaining variation in the response variables measured. In analyses based on Hedges' *d* we found that the mean |*d*| across 21 meta-analyses was 0.63. This is slightly larger than an intermediate effect size of *d*=0.5 (Cohen 1988). Although it may be misleading to convert from *d* to r^2 and then calculate *r*_ (Gurevitch, personal communication), Cohen (1988, p 26) notes that his definition of an intermediate value of *d*=0.5 is such that "6% of the variance is 'accounted for' by populational membership".

Biology differs from other natural sciences by dealing with living organisms that are affected by innumerable biotic and abiotic factors. It is therefore no surprise that the amount of variance accounted for by any single factor in ecology and evolution is relatively small. This contrasts strongly with studies in physics and chemistry, where the amount of variance explained is typically very large. Of course, even a minute effect size may be biologically important, in particular when discussing evolutionary issues. Small effects may become greatly magnified when a persistent pattern occurs across many generations.

A paired comparison showed no clear increase in effect size when using an experimental approach over simple observations. The power to detect a medium-sized difference was, however, only 25%. Interestingly though, the two kinds of estimates of effect size were strongly positively correlated. Similarly, Gontard-Danek and Møller (1999) found only a slightly larger effect size in experimental compared to observational studies of sexual selection, even when they were paired for the same species. This seems unlikely to reflect the situation at large. It is likely that some systems are more tractable experimentally, simply for logistic or other practical reasons, and such systems will contribute disproportionately to the body of knowledge in a field (see also Thornhill et al. 1999). Given our small sample size, further studies of the increase in variation that can be explained when using an experimental approach would be of general interest.

The fail-safe number was strongly positively correlated with the amount of variance explained or the magnitude of the effect size in different meta-analyses (Fig. 3). This is intuitively expected, since a larger number of null results are needed to nullify a larger mean effect. This

relationship is particularly strong though, given that sample sizes varied considerably among studies. The considerable amount of variance in fail-safe number for a given effect size (Fig. 3) clearly supports the observation that a large fail-safe number does not automatically follow from a large effect size. Only when a consistent effect size has been found in a large number of studies will the fail-safe number be large.

We found weak evidence for significant differences in mean effect size among fields of biology, although the 95% confidence intervals overlapped (Fig. 4). Studies with a physiological content tended to have larger Hedges' *d* than ecological ones, which in turn had larger values than evolutionary studies. Meta-analyses examining the role of fluctuating asymmetry (or other measures of developmental instability) in ecological and evolutionary questions yielded mean effect sizes very similar to those obtained in other evolutionary studies. Thus, the relatively small effect sizes reported in meta-analyses of asymmetry do not differ significantly from those reported in other fields. For example, analyzed at the meta-analysis level, the weighted mean effect size for asymmetry studies was identical to that for other evolutionary studies at $|r|=0.203$. Although it has been suggested that studies of asymmetry generally have small effect sizes and therefore are of little biological significance (Houle 1998), and that publication bias has inflated estimates of effect size in studies of asymmetry because of a negative relationship between sample size and effect size (Palmer 1999) and a temporal decline in effect size with date of publication (Simmons et al 1999), these appear to be general trends in biological studies (e.g. Alatalo et al 1997; Møller and Alatalo 1999; Poulin 2000; review in Jennions and Møller 2002). Clearly the mean effect size in studies of asymmetry is not different from the mean effect size in all meta-analyses. The approach that we have adopted here allows scientists to judge the mean effect size in a given field without biasing their statements with personal opinion.

The amount of variance explained by a given factor of interest depends on the extent to which confounding variables are controlled for, either experimentally or statistically. For example, the estimated effect of a treatment will be far stronger when an experiment is designed to control for a significant covariate or analyzed using ANCOVA than would be the case if this covariate went unmeasured and analysis was based on a two-sample *t*-test. Even so, in the absence of more detailed information or pilot studies, the mean effect sizes we have reported here can be used by researchers to estimate the sample sizes they will need to detect average effects with a given statistical power. To date, most researchers have assumed a medium affect of $r=0.3$ as defined by Cohen (1988). The 95% confidence intervals for mean effects calculated as *r* all fell between $|r|=0.14$ and 0.25. These should therefore be used to indicate a reasonable range of sample sizes. With 80% statistical power these correspond to sample sizes of 122 and 396. It is worth noting that very few statistical tests in evolutionary biology (and especially behavioral ecology) are based on sample sizes this large, especially when calculating correlations. That was clearly also the case in most of the studies included in the meta-analyses investigated here. Thus, overall significant effects in many meta-analyses arose as a consequence of many studies showing an effect in a particular direction, rather than scientists using a large sample size to have a high level of power and thereby finding statistically significant results. Failure to reject the null hypothesis should therefore be interpreted with far greater caution, and a heightened appreciation of what an 'average' strength relationship in evolutionary biology really is. The 95% confidence intervals for mean Hedges' *d* fell between 0.48 and 0.82. Although non-identical, Hedges' *d* and Cohen's *d* are closely related estimates of effect size. To detect a Cohen's *d* of 0.48 or 0.82 requires a sample size of 24 and 69 per group respectively (Cohen 1988). Again, many studies in biology that compare two sample populations tend to be based on smaller sample sizes.

In conclusion, a meta-analysis of meta-analyses in ecology and evolution revealed small to intermediate effect sizes (sensu Cohen 1988). The amount of variance explained decreased from physiology over ecology to evolution. These findings suggest that biological studies, even experimental ones, will often only explain a very small amount of variance. The merits of different avenues of research should be evaluated in the light of these findings, including improved designs of observational and experimental studies, sufficient sample sizes to obtain a reasonable power, and more widespread use of meta-analysis.

Acknowledgements J. Shykoff kindly discussed issues of meta-analysis and the general idea behind performing the present study. Göran Arnqvist, Michael Brett, Isabelle Côté, Peter Curtis, Mark Forbes, Peter Hamback, Nick Jonsson, Julia Koricheva, Dean McCurdy, Fiorenza Micheli, Iago Mosqueira, Robert Poulin, Howie Riessen, Michael Rosenberg, Gina Schalk, Xianzhong Wang and Peter van Zandt kindly provided unpublished information. We thank Jessica Gurevitch and two anonymous reviewers for their constructive comments.

Appendix

The 43 meta-analyses included in the present study. Coding is presented in parenthesis as ecological (EC), evolutionary (EV) and phsyiological (P); and dealing with asymmetry (FA) or not (N).

Arnqvist G, Nilsson T (2000) The evolution of polyandry: multiple mating and female fitness in insects. Anim Behav 60:145–164 (EV, N)

Arnqvist G, Rowe L, Krupa JJ, Sih A (1996) Assortative mating by size: a meta-analysis of mating patterns in water striders. Evol Ecol 10:265–284 (EV, N)

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

Brett MT, Goldman G (1996) A meta-analysis of the freshwater trophic cascade. Proc Natl Acad Sci USA 93: 7723–7726 (EC, N)

Cadée N, Møller AP (2000) On the relative sensitivity of trait size and asymmetry to environmental stress. In: Cadée N Ecological aspects of stress resistance in the barn swallow *Hirundo*

rustica. PhD thesis, Laboratoire d'Ecologie, Université Pierre et Marie Curie, Paris, France, pp 52–89 (EV, FA)
Côté IM, Poulin R (1995) Parasitism and group size in social animals: a meta-analysis. Behav Ecol 6:159–165 (EV, N)
Côté IM, Sutherland WJ (1997) The effectiveness of removing predators to protect bird populations. Conserv Biol 11:395–405 (EC, N)
Curtis PS (1996) A meta-analysis of leaf gas exchange and nitrogen in trees grown under elevated carbon dioxide. Plant Cell Environ 19:127–137 (P, N)
Curtis PS, Wang X (1998) A meta-analysis of elevated CO_2 effects on woody plant mass, form, and physiology. Oecologia 113: 299–313 (P, N)
Fernandez-Duque E, Valeggia C (1994) Meta-analysis: a valuable tool in conservation research. Conserv Biol 8: 555–561 (EC, N)
Fiske P, Rintamäki P, Karvonen E (1998) Mating success in lekking males: a meta-analysis. Behav Ecol 9: 328–338 (EV, N)
Gontard-Danek M-C, Møller AP (1999) The strength of sexual selection: a meta-analysis of bird studies. Behav Ecol 10:476–486 (EV, N)
Gurevitch J, Morrow LL, Wallace A, Walsh JS (1992) A meta-analysis of competition in field experiments. Am Nat 140: 539–572 (EC, N)
Hamilton WJ, Poulin R (1997) The Hamilton and Zuk hypothesis: a meta analytic approach. Behaviour 134: 299–320 (EV, N)
Harper DGC (1999) Feather mites, pectoral muscle condition, wing length and plumage coloration of passerines. Anim Behav 58:553–562 (EV, N)
Järvinen A (1991) A meta-analytic study of the effects of female age on laying-date and clutch-size in the Great Tit *Parus major* and the Pied Flycatcher *Ficedula hypoleuca*. Ibis 133:62–67 (EC, N)
Jennions MJ, Møller AP, Petrie M (2001) Sexually selected traits and adult survival: a meta-analysis of the phenotypic relationship. Q Rev Biol 76:3–36 (EV, N)
Koricheva J (2001) Meta-analysis of sources of variation in fitness costs of plant antiherbivore defenses. Ecology (in press) (P, N)
Koricheva J, Larsson S, Haukioja E (1998a) Insect performance on experimentally stressed woody plants: a meta-analysis. Annu Rev Entomol 43:195–216 (EC, N)
Koricheva J, Larsson S, Haukioja E, Keinänen M (1998b) Regulation of woody plant secondary metabolism by resource availability: hypothesis testing by means of meta-analysis. Oikos 83: 212–226 (P, N)
Leung B, Forbes MR (1996) Fluctuating asymmetry in relation to stress and fitness: effects of trait type as revealed by meta-analysis. Ecoscience 3:400–413 (EV, FA)
Møller AP (1999) Asymmetry as a predictor of growth, fecundity and survival. Ecol Lett 2:149–156 (EV, FA)
Møller AP (2000) Developmental stability and pollination. Oecologia 123:149–157 (EV, FA)
Møller AP, Alatalo RV (1999) Good genes effects in sexual selection. Proc R Soc Lond Ser B 266:85–91 (EV, N)
Møller AP, Ninni P (1998) Sperm competition and sexual selection: a meta-analysis of paternity studies of birds. Behav Ecol Sociobiol 43:345–358 (EV, N)
Møller, AP, Shykoff JA (1999) Developmental stability in plants: Patterns and causes. Int J Plant Sci 160:S135-S146 (EV, FA)
Møller AP, Christe P, Erritzøe J, Mavarez J (1998) Condition, disease and immune defence. Oikos 83:301–306 (EV, N)
Møller AP, Christe P, Lux E (1999) Parasite-mediated sexual selection: Effects of parasites and host immune function. Q Rev Biol 74:3–20 (EV, N)
Poulin R (1994) Meta-analysis of parasite-induced behavioural changes. Anim Behav 48:137–146 (EV, N)
Poulin R (1996) Sexual inequalities in helminth infections: a cost of being a male? Am Nat 147:287–295 (EV, N)
Poulin R (2000) Variation in the intraspecific relationship between fish length and intensity of parasitic infection: biological and statistical causes. J Fish Biol 56:123–137 (EV, N)
Riessen HP (1999) Predator-induced life history shifts in *Daphnia*: a synthesis of studies using meta-analysis. Can J Fish Aquat Sci 56:123–137 (EV, N)
Schalk G, Forbes MR (1997) Male biases in parasitism of mammals: effects of study type, host age, and parasite taxon. Oikos 78:67–74 (EV, N)
Sokolovska N, Rowe L, Johansson F (2000) Fitness and body size in mature odonates. Ecol Entomol 25:239–248 (EV, N)
Thornhill R, Møller AP (1998) The relative importance of size and asymmetry in sexual selection. Behav Ecol 9:546–551 (EV, N)
Thornhill R, Møller AP, Gangestad SW (1999) The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. Am Nat 154:234–241 (EV, FA)
Tonhasca AJ, Byrne DN (1994) The effects of crop diversification on herbivorous insects: a meta-analytic approach. Ecol Entomol 19:239–244 (EC, N)
Van Zandt PA, Mopper S (1998) A meta-analysis of adaptive deme formation in phytophagous insect populations. Am Nat 152:595–604 (EV, N)
VanderWerf E (1992) Lack's clutch size hypothesis: an examination of the evidence using meta-analysis. Ecology 73:1699–1705 (EV, N)
Vøllestad LA, Hindar K, Møller AP (1999) A meta-analysis of fluctuating asymmetry in relation to heterozygosity. Heredity 83:206–218 (EV, FA)
Wang X, Curtis PS (2001) A meta-analytical test of elevated CO2 effects on plant respiration. Plant Ecol (in press) (P, N)
Wooster D (1994) Predator impacts on stream benthic prey. Oecologia 99:7–15 (EC, N)
Xiong S, Nilsson C (1999) The effect of plant litter on vegetation: a meta-analysis. J Ecol 87:984–994 (EC, N)

References

Alatalo RV, Mappes J, Elgar MA (1997) Heritabilities and paradigm shifts. Nature 385:402–403
Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manage 64:912–923
Arnqvist G, Nilsson T (2000) The evolution of polyandry: multiple mating and female fitness in insects. Anim Behav 60:145–164
Arnqvist G, Rowe L, Krupa JJ, Sih A (1996) Assortative mating by size: a meta-analysis of mating patterns in water striders. Evol Ecol 10:265–284
Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. Biometrics 50:1088–1101
Berlin JA, Begg CB, Louis TA (1989) An assessment of publication bias using a sample of published clinical tests. J Am Stat Assoc 84:381–392
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale, N.J.
Cooper H, Hedges LV (eds)(1994) The handbook of research synthesis. Russell Sage, New York
Curtis PS (1996) A meta-analysis of leaf gas exchange and nitrogen in trees grown under elevated carbon dioxide. Plant Cell Environ 19:127–137
Davis WJ (1993) Contamination of coastal versus ocean surface waters: a brief meta-analysis. Mar Pollut Bull 26:128–134
Dear KBG, Begg CB (1992) An approach for assessing publication bias prior to performing a meta-analysis. Stat Sci 7:237–245
Gontard-Danek M-C, Møller AP (1999) The strength of sexual selection: a meta-analysis of bird studies. Behav Ecol 10:476–486
Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. Ecology 80:1142–1149
Haldane JBS (1949) Suggestions as to quantitative measurement of rates of evolution. Evolution 3:51–56
Harper DGC (1999) Feather mites, pectoral muscle condition, wing length and plumage coloration of passerines. Anim Behav 58:553–562
Hedges LV (1992). Meta-analysis. J Educ Stat 17:279-296
Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press, San Diego, Calif.
Hendry P, Kinnison MT (1999) The pace of modern life: measuring rates of contemporary microevolution. Evolution 53:1637–1653

Houle D (1998) High enthusiams and low R-squared. Evolution 52:1872–1876
Hunter JE, Schmidt FL (1990) Methods of meta-analysis: correcting error and bias in research findings. Sage, Beverly Hills, Calif.
Jennions MD, Møller AP (2002) Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proc R Soc Lond Ser B 269:43–48
Light RJ, Pillemer DB (1984) Summing up: The science of reviewing research. Harvard University Press, Cambridge, Mass.
Maynard Smith J (1978) Optimization theory in evolution. Annu Rev Ecol Syst 9:31–56
Mengersen KL, Tweedie RL, Biggerstaff BJ (1995) The impact of method choice in meta-analysis. Aust J Stat 7:19–44
Møller AP, Alatalo RV (1999) Good genes effects in sexual selection. Proc R Soc Lond Ser B 266:85–91
Møller AP, Jennions MD (2001) Testing and adjusting for publication bias. Trends Ecol Evol 16:580–586
Osenberg CW, Sarnelle O, Cooper SD (1997) Effect size in ecological experiments: the application of biological models in meta-analysis. Am Nat 150: 798–812
Palmer AR (1999) Detecting publication bias in meta-analysis: a case study of fluctuating asymmetry and sexual selection. Am Nat 154:220–233
Palmer AR (2000) Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. Annu Rev Ecol Syst 31:441–480
Poulin R (2000) Manipulation of host behaviour by parasites: a weakening paradigm? Proc R Soc Lond Ser B 267:787–792
Ridley M (1993) Evolution. Blackwell, Oxford
Rosenberg MS, Adams DC, Gurevitch J (2000) MetaWin: Statistical Software for Meta-Analysis. Version 2.0. Sinauer, Sunderland, Mass.
Rosenthal R (1991) Meta-analytic procedures for social research. Sage, New York
Rosenthal R (1994) Parametric measures of effect size. In H Cooper, LV Hedges (eds) The handbook of research synthesis. Sage, New York, pp 231–244
Simmons LW, Tomkins JL, Kotiaho JS, Hunt J (1999) Fluctuating paradigm. Proc R Soc Lond Ser B 266:593–595
Simpson GG (1944) Tempo and mode in evolution. Columbia University Press, New York
Simpson GG (1949) Rates of evolution in animals. In: Jepsen GL, Simpson GG, Mayr E (eds) Genetics, paleontology, and evolution. Princeton University Press, Princeton, N.Y. pp 205–228
Thompson SG (1993) Controversies in meta-analysis: the case of trials of serum cholesterol reduction. Stat Methods Med Res 2:173–192
Thornhill R, Møller AP, Gangestad S (1999) The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. Am Nat 154:234–241
Vandenbroucke JP (1988) Passive smoking and lung cancer: a publication bias? Br Med J 296:91–392
Warwick RM, Clarke KR (1993) Comparing the severity of disturbance: A meta-analysis of marine macrobenthic community data. Mar Ecol Progr Ser 92:221–231