Meta-analysis of magnitudes, differences, and variation in evolutionary parameters

Michael B. Morrissey

May 31, 2016

School of Biology, University of St Andrews

contact
email: michael.morrissey@st-andrews.ac.uk
phone: +44 (0) 1334 463738
fax: +44 (0) 1334 463366
post: Dyers Brae House
School of Biology, University of St Andrews
St Andrews, Fife, UK, KY16 9TH

Keywords: meta-analysis, synthesis, natural selection, reaction norms

Abstract

Meta-analysis is increasingly used to synthesise major patterns in the large literatures within ecology and evolution. Meta-analytic methods that do not account for the process of observing data, which we may refer to as 'informal meta-analyses', may have undesirable properties. In some cases, informal meta-analyses may produce results that are unbiased, but do not necessarily make the best possible use of available data. In other cases, unbiased statistical noise in individual reports in the literature can potentially be converted into severe systematic biases in informal meta-analyses. I first present a general description of how failure to account for noise in individual inferences should be expected to lead to biases in some kinds of meta-analysis. In particular, informal meta-analyses of quantities that reflect the dispersion of parameters in nature, for example, the mean absolute value of a quantity, are likely to be generally highly misleading. I then re-analyse three previously published informal meta-analyses, where key inferences were of aspects of the dispersion of values in nature, for example, the mean absolute value of selection gradients. Major biological conclusions in each original informal meta-analysis closely match those that could arise as artefacts due to statistical noise. I present alternative mixed model-based analyses that are specifically tailored to each situation, but where all analyses may be implemented with widely available open-source software. In each example meta-re-analysis, major conclusions change substantially.

1 Introduction

Many questions in ecology and evolution concern the distribution of effects across space, time, taxa, and ecological conditions. Consequently, synthetic works have a critical role to play in organising the general knowledge that accumulates in the vast literatures within ecology and evolution. Recently, meta-analytical approaches have become increasingly popular for describing accumulated results (Nakagawa and Poulin, 2012).

Meta-analyses are studies that employ a quantitative approach to draw robust conclusions about natural phenomena, by drawing on all available and appropriate estimates, typically as reported in the primary scientific literature. This is an intentionally inclusive definition, appealing to the motivation, conception, and likely perceived comprehensiveness and general validity, of meta-analytic exercises. This definition is consistent with the original (Glass, 1976) and subsequent (Gurevitch and Hedges, 1999; Nakagawa and Santos, 2012; O'Rourke, 2007) uses of the term. Within exercises conducted in the meta-analytic spirit, a range of approaches exists. 'Informal meta-analysis', as I will refer to some studies conducted in the meta-analytic spirit, make inferences about phenomena in nature (for example, the effect of an environmental perturbation on some aspect of a species' biology, or the strength of natural selection) by reporting summary statistics of the distribution of estimated values in a meta-dataset (i.e., a database constructed from the available literature). While the motivation, and typically the perceived validity, of such studies falls entirely within the domain of the meta-analytic enterprise, some authors object to their characterisation as meta-analyses, preferring instead to categorise as meta-analyses only those studies that use specific statistical methods that are deemed to be meta-analytical (Koricheva and Gurevitch 2013*a*, page 8; Vetter et al. 2013). More 'formal meta-analyses' will generally apply some system for accounting for the varying precision or quality of individual elements of a meta-database. However, it seems undesirable to place arbitrary limits on what such methods should be.

Some meta-analyses will investigate average effects, i.e., means of distributions of quantities, or factors that influence the mean, such as covariates or "moderator variables" (Nakagawa and Santos, 2012). For example, a meta-analysis in a conservation context may seek to determine

whether some environmental condition has a negative impact on some aspect of an organism's biology. Sometimes, the key questions of interest pertain to higher-order aspects of the distributions of effects. We may be interested in the *average magnitudes*, or average absolute values, of some phenomena, rather than the *average values*. For example, the directionality of many phenomena, such as the form of natural selection, is either arbitrary in general (selection of development rate vs. development time), or is arbitrary at the level of meta-data. We might therefore be interested in the variance or standard deviation of effects, the averages of absolute values, the average magnitude of differences between treatments, or other aspects of the variation in effects.

Statistical noise, or sampling error, generates variation in *estimated parameter values*, over and above any true variation in those parameter values. Consequently, informal meta-analyses of some types of parameters will generally mistake unbiased statistical noise at the level of individual parameter estimates for biologically interesting variation at the level of meta-datasets. In general, informal meta-analytic inference of the means of natural phenomena will be unbiased by sampling error (this assertion conflicts with a recent survey of the topic Koricheva and Gurevitch 2013*b*; see further formal treatment below). Other quantities, such as average magnitudes (i.e., mean absolute values), will be upwardly biased in informal meta-analyses. For example, variation in *estimated* selection gradients in temporally replicated studies can be erroneously interpreted as evidence for pervasive variation in natural selection, if sampling error is not taken into account (Morrissey and Hadfield, 2012; Siepielski et al., 2009). Additionally, complexities in the observation process in individual studies, over and above pure statistical noise, can also generate spurious, but superficially biologically interesting and convincing, results in meta-analyses. For example, the inclusion of studies conducted at different scales can generate serious spurious meta-analytical patterns in synthetic studies of species richness-productivity relationships (Whittaker, 2010).

Here I first analyse some simple models of meta-analyses. This clarifies what types of informal meta-analyses may be, or may not be, biased by statistical noise in individual studies. I then conduct a simulation study of the performance of three different approaches to meta-analysis, specifically focusing on cases where interest is not directly in the quantities that are

reported in the literature, but rather in some derived value. For example, a derived value may be the absolute value (e.g., magnitude) of some quantity, when what is actually reported in the literature is the quantity itself, not the absolute value. I suggest a general approach of modelling distributions of quantities in the literature as they are reported, and then subsequently deriving different quantities that may be of interest. I then re-analyse three important informal meta-analyses. In each instance, I first present simple arguments showing why the main results in each of three different informal meta-analyses are inevitably and strongly influenced by sampling error. I discuss, in each situation, how white noise at the level of individual studies is converted to biases by informal meta-analytic procedures. For each study, I present alternative model-based versions of the key analyses. In each case, major results change substantially.

2 Statistical noise and bias in meta-analysis: a model

In this section, I consider a very simple model of a meta-analysis. This allows both analytical and simulation results to be presented to show different situations where meta-analyses might be unbiased or biased.

2.1 Model structure

I assume that N studies exist, each reporting a single estimate of some quantity, x. Each estimate of x will be denoted $\hat{x}_i$; the "hat" symbol indicates that we are dealing with an estimate, not a known quantity, and i indexes the estimates from the N studies. I assume that each available value of $\hat{x}_i$ is obtained by some method (which may differ among the N studies) that is unbiased. Formally, "unbiased" means that for each estimate,

$$E[\hat{x}_i] - x_i = 0. \tag{1}$$

Of course, each estimate is not the true value, i.e., we do not require that $\hat{x}_i = x_i$. Rather, across many estimates, $\hat{x}_i$, we require that the true value is not, on average either over- or under-estimated. Many statistical procedures in common use, when used correctly, provide unbiased estimates of natural phenomena. For example, $\hat{x}$ values could be estimates of the

mean, or regression slopes from least-squares analysis.

True values of the parameter of interest, i.e., of the x_i, are assumed to come from some distribution. For simplicity, I model that true values as normally distributed. Formally, we can write this as

$$x_i \sim N\left(\mu_x, \sigma_x^2\right), \tag{2}$$

which simply states that each (in practice, unknown) true value is drawn from a normal distribution with some mean (μ_x) and variance (σ_x^2). Features of the distribution of true values of x that may be of interest in a meta-analysis could be the mean (μ_x), the variance (σ_x^2), or some other property of the distribution of x, such as the mean absolute value $E[|x|]$.

I also assume that each estimate is associated with information about its uncertainty. We cannot know the true values, x_i, associated which each estimate $\hat{x}_i$ in a meta-database. Rather, each $\hat{x}_i$ value will be drawn from some distribution defined by the true value, x, and its measurement error. For simplicity, I assume that the distributions of measurement errors are normal, such that

$$\hat{x}_i = x_i + e_i, \tag{3a}$$

$$e_i \sim N\left(0, \sigma^2(m)_i\right), \tag{3b}$$

which simply states that each estimate is drawn from a normal distribution around the true value for that study, and the "noise" in the $\hat{x}_i$ values around the x_i values is defined by each estimate's sampling variance, $\sigma^2(m)_i$ (which is the square of the standard error). Conclusions drawn assuming normal sampling error should be quite generally informative: for example, the sampling distribution of a mean (if x_i values are the means of some quantity in each study) is t-distributed, but this distribution approaches a normal distribution quite rapidly with increasing sample size.

2.2 Meta-analysis of the mean

We may be interested in the mean of some quantity in nature. In our model, this is μ_x. For example, our x_i values may be differences in bird singing volume between two habitats (e.g.,

natural vs. urban), and we may be interested in the overall mean difference, μ_x. We might estimate the overall mean by

$$\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_i, \tag{4}$$

i.e., our estimator of μ_x, $\hat{\mu}_x$, may simply be the average of all available estimates.

A number of sources on meta-analysis place emphasis on the need to weight results from individual studies in some way determined by their sampling variance (e.g., Arnqvist and Wooster 1995; Koricheva et al. 2013; Vetter et al. 2013). These views represent cautions against analyses such as that represented by equation 4. For example, *Handbook of Meta-analysis in Ecology and Evolution* chapter 7 page 81, Koricheva and Gurevitch (2013*b*) state that:

> ...it is essential to be able to derive a variance [meaning $\sigma^2(e)_i$ in the model here] for the metric obtained in each study [for each $\hat{x}_i$], and to use these to weight the effect sizes in the meta-analysis. Unweighted analyses produce biased estimates of overall effects [e.g., of quantities such as μ_x].

Formally, this view contends that

$$E[\hat{\mu}_x] - \mu_x \neq 0$$

when $\hat{\mu}_x$ is that obtained by the informal meta-analysis method in equation 4. Of course we never know μ_x, and so we never know whether our estimate, $\hat{\mu}_x$, is too large or small in any given case. However, we can use statistical theory and/or simulation to determine whether a given meta-analytic procedure, such as that in equation 4, would on average give too high or too low an estimate, if applied over many meta-analyses. Equation 3 states that the mean of sampling errors is zero (this is just a corollary of the assumption reports of $\hat{x}$ in the literature are unbiased). In general the expectation of a sum is equal to the sum of expectations[1]: $E[A+B] = E[A]+E[B]$. For our possible meta-analysis in equation 4, the mean of true values and the mean of sampling errors would correspond to $E[A]$ and $E[B]$. These are defined as μ_x (in equation 2) and zero (in equation 3b), respectively. So, $E[x+e] = E[x]+E[e] = \mu_x+0 = \mu_x$.

[1] $E[A+B]$ can be written as all possible values of the sum of A and B, weighted by the probability density of each possible set of values of A and B: $E[A+B] = \int_A \int_B (A+B) f(A,B) dB dA$, where $f(A,B)$ is an arbitrary joint probability function of A and B. Using the summation/subtraction rule: $E[A+B] = \int_A \int_B A f(A,B) dB dA + \int_A \int_B B f(A,B) dB dA$. The expression simplifies: $E[A+B] = \int_A A f(A) dA + \int_B B f(B) dB$. Since $E[X] = \int X F(X) dX$, $E[A+B] = E[A]+$ E[B].

Therefore, provided that each $\hat{x}_i$ is an unbiased estimate of x_i, then the mean of $\hat{x}_i$ values is an unbiased estimator of μ_x. This proves that an average of unbiased estimates of x, i.e., of $\hat{x}_i$ values, is an unbiased estimator of their means, even if no formal meta-analysis is implemented.

Just because a simple summary statistic of values in a meta-database is not biased does not necessarily mean that it is the best analytical approach. In general, different studies will have different sampling variances. Those $\hat{x}$ values with the smallest sampling variances contain the most reliable information about the true distribution of x. Weighting schemes for calculating meta-analytic estimates of quantities such as μ_x (reviewed in Koricheva et al. 2013) have been developed to minimise the sampling variance of meta-analytic quantities, i.e., to make them as precise as possible, and not to reduce bias. When information about statistical uncertainty is available (e.g., when standard errors are reported), such approaches should be used. However, in the absence of standard errors, or when they are inconsistently reported, it is possible that an informal, summary statistic-based, meta-analysis such as that represented by equation 4 can be highly precise (potentially more precise than a formal meta-analysis that can only use a restricted database of estimates with standard errors) and unbiased.

2.3 Meta-analysis of the mean absolute value (i.e., the average magnitude)

However, there is no guarantee that any particular informal meta-analysis will be unbiased. In this section I consider that a meta-analysis may seek to determine, not the mean of x, but the average magnitude of x. These may seem like very similar problems, but we will see that meta-analyses of these different parameters involve very different considerations.

For simplicity, assume that all estimates of x have the same standard error, and therefore that all values of $\sigma^2(e)_i$ are equal. In our model, both true values and sampling errors are normal, and so the distribution of estimates is also normal. Situations where the mean magnitude will be of interest will often be when the mean is close to zero, such that both positive and negative values occur; so an simple instructive case to consider will be the situation when $\mu_x = 0$. The mean absolute value of a centred normally-distributed variable is the mean of a χ distribution with one degree of freedom, times the standard deviation of that variable (this arises simply

from the definition of the χ distribution). The mean of a χ distribution is $\sqrt{2}\frac{\Gamma((k-1)/2)}{\Gamma(k/2)}$, where $\Gamma()$ represents the gamma function. We are interested in the situation where $k = 1$, and so using $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ we obtain

$$E[|x|] = \sqrt{\frac{2}{\pi}}\sigma(x) \tag{5}$$

when $\mu_x = 0$. This equation for the mean absolute value of a centred normal variable allows us to obtain an expression for bias in a summary statistic-based meta-analysis of mean absolute values. If we were to estimate mean absolute value by

$$\hat{\mu}_{|x|} = \frac{1}{N}\sum_{i=1}^{N}|\hat{x}_i|,$$

then the expected value of this estimator would be

$$\sqrt{\frac{2}{\pi}}\sqrt{\sigma^2(x) + \sigma^2(m)}.$$

$\sqrt{\sigma^2(x) + \sigma^2(e)}$ is the standard deviation of estimates of x, assuming errors to be independent of true values. In contrast, the mean absolute value of true values of x would be

$$\sqrt{\frac{2}{\pi}}\sigma(x).$$

From the definition of bias, we can obtain the bias in the informal meta-analysis of mean absolute values as

$$\begin{aligned} E[\hat{\mu}_{|x|}] - E[|x|] &= \sqrt{\frac{2}{\pi}}\sqrt{\sigma^2(x) + \sigma^2(m)} - \sqrt{\frac{2}{\pi}}\sigma(x) \\ &= \sqrt{\frac{2}{\pi}}\left(\sqrt{\sigma^2(x) + \sigma^2(m)} - \sqrt{\sigma^2(x)}.\right) \end{aligned} \tag{6}$$

If there is any sampling error in estimates of x, then $\sqrt{\sigma^2(x) + \sigma^2(e)}$ will be greater than $\sqrt{\sigma^2(x)}$, and the summary statistic-based meta-analysis of mean absolute value will be upwardly biased.

3 Analytical options for meta-analysis: a small simulation study

Here, I explore the results of three possible meta-analytic procedures for inference of means and mean absolute values, i.e., average magnitudes, of arbitrary quantities. The first method is an informal, summary statistic-based meta-analysis. The second option is to derive sampling variances of any derived quantities in a meta-database, for use with established meta-analytic procedures. This is the standard approach in meta-analysis, though transformation is often not required. I refer to this as the "transform-then-analyse" approach. The third option is to apply meta-analytic mixed model analysis to estimate parameters of the distribution of x (i.e., the quantities in the literature as they are reported, even if some transformation of x, say the absolute value, is ultimately of interest), accounting for sampling error in individual $\hat{x}_i$ estimates, and then to derive the desired quantity of interest (e.g., $E[|x|]$). I refer to this as the "analyse-then-transform" approach. This last approach has previously been used as an alternative to summary statistic-based informal meta-analysis (see Morrissey and Hadfield 2012's re-analysis of temporal variation in selection as first reported on by Siepielski et al. 2009), but it has yet not been explored as a general approach to meta-analysis.

3.1 Simulation scheme

For each replicate simulation, I simulated a meta-database of 50 studies. Each study had one associated value of $\hat{x}_i$ and an associated standard error, $\sigma^2(m)_i$. The $\hat{x}_i$ values were drawn from a normal distribution according to $\hat{x}_i \sim N(\mu_x, \sigma^2(m)_i)$, and the true values of x were simulated according to $x_i \sim N(\mu_x, \sigma^2(x))$. This closely follows the model that was investigated analytically, above. I simulated all combinations of values of μ_x of 0 and 0.25, and a range of values of $\sigma^2(x)$ between 0.01 and 1.0. Furthermore, for all combinations of values, I simulated two different average magnitudes of statistical noise. Each x_i value's associated value of $\sigma^2(m)_i$ was drawn from a gamma distribution with mean and standard deviation of either 0.25 or 0.5. This is merely a convenient way of ensuring that some estimates within each simulated meta-analysis are more precise than others (while none is absolutely perfect), and also of simulating meta-analyses that contend with different overall levels of statistical noise. For each

combination of true mean and variance of x, and of statistical noise, I simulated 1000 replicate meta-analyses.

The true overall mean of x, i.e. μ_x, is simply one of the parameters of the simulation. However, the true value mean absolute value of x is determined both by μ_x and by $\sigma^2(x)$. As such, the true value of $E[|x|]$ in each study is defined by a folded normal distribution

$$\bar{\mu_{|x|}} = \sqrt{\frac{2}{\pi}}\sigma(x)e^{-\mu_x^2/2\sigma^2(x)} + \mu_x(1 - 2\Phi(\frac{-\mu_x}{\sigma(x)})), \tag{7}$$

which is simply the mean of a normal distribution defined by μ_x and $\sigma^2(x)$, folded about the origin.

For each simulation, I implemented the informal meta-analyses of the mean and mean absolute value by calculating the mean of the simulated $\hat{x}_i$ values, and the mean of their absolute values. In order to implement the 'transform-then-analyse' meta-analysis, I had to first obtain the sampling variance of the transformed values of $\hat{x}_i$, i.e., the sampling variance of $|\hat{x}_i|$. This is defined by the variance of a folded normal distribution, for each $\hat{x}_i$ and and its corresponding sampling variance $\sigma^2(m)_i$

$$\sigma^2(m)_{|\hat{x}_i|} = \hat{x}_i^2 + \sigma^2(m)_i - \left(\sqrt{\frac{2}{\pi}}\sigma(m)e^{-\hat{x}^2/2\sigma^2(m)_i} + \hat{x}_i(1 - 2\Phi(\frac{-\hat{x}_i}{\sigma(m)_i}))\right)^2. \tag{8}$$

I then applied a mixed-model based meta-analysis of the $|\hat{x}_i|$ values and their derived sampling variances. A mixed model meta-analysis is a generalisation of various weighting schemes that exist in the meta-analysis literature. The mixed model took the form

$$y_i = \mu_y + m_i + e_i, \tag{9}$$

where y_i are the data in the meta-analytic database; in the 'transform-then-analyse' procedure, the y_is are the $|\hat{x}_i|$ values. μ_y is the model intercept, which is the meta-analytic estimator of the mean of whatever the y_i values are. m_i are the measurement errors for each value of y_i. Of course we cannot know these errors in each case, but the model integrates over the possible values that the m_i can take, using the information available about their sampling

variances. This is accomplished by defining the measurement errors to come from a distribution $m_i \sim N(0, \sigma^2(m)_i)$, where the sampling variances $\sigma^2(m)_i$ are appropriate to whatever the y_i are; in the case of the simulated 'transform-then-analyse' meta-analyses, the $\sigma^2(m)_i$ values associated with the $|\hat{x}_i|$ values are those given by equation 8. Finally, the residuals, i.e., the e_i values are modelled according to $e_i \sim N(0, \sigma^2(e))$, where $\sigma^2(e)$ is estimated by the mixed model. $\sigma^2(e)$ is thus the meta-analytic estimator of the variance of x, i.e., of $\sigma^2(x)$ in the notation used in the analytical sections, above.

Finally, the 'analyse-then-transform' meta-analysis was simulated using a mixed model of the form described by equation9, except the $\hat{x}_i$ values were used for the y_i, along with their associated sampling variances (the simulated standard errors, squared). This provided meta-analytic estimates of the simulated μ_x and $\sigma^2(x)$ values (i.e., the μ_y and $\sigma^2(e)$ values estimated from the mixed model). These estimates were then used to obtain estimated mean absolute values, using the expression for the mean of a folded normal distribution (equation 7). I fitted all meta-analytic mixed models using the *rma()* function from the R package METAFOR (Viechtbauer, 2010).

4 Simulation results, and conclusions from analytical models and simulations

As suggested by theory, all three meta-analytic approaches yielded unbiased results of the overall means, and are not considered further. Also as expected from analytical results (equation 6), naive summary statistic-based meta-analysis of mean absolute values are upwardly biased, across a range of parameters (figure 1). Simulation results support various features of the analytical expression for bias (equation 6): the bias is greatest when sampling variance is high, and especially when sampling variances are high relative to true variances. While the theoretical analysis did not deal with situations where the true mean is non-zero[2], the simulations give

[2]Expressions for bias in the mean absolute value when the mean is non-zero can be written down; however, I was unable to make them simple enough to be generally informative. Expressions for bias in informal meta-analysis of mean absolute values can be constructed either using folded normal distributions or the non-central χ distribution. In both cases, the expressions involve complicated functions, the parameterisation using the folded normal involves the error function, and the parameterisation using the non-central χ distribution requires generalised Laguerre polynomials; neither is conducive to useful simplifications.

fairly intuitive results. When the true mean is not zero, mean absolute values are less biased, in informal meta-analyses.

For the range of parameters investigated, the standard 'transform-then-analyse' formal meta-analytic approach was consistently biased. The bias was intermediate between the naive meta-analysis and the 'analyse-then-transform meta-analysis'. The bias in this formal approach to meta-analysis arises because the model for sampling error in the random effects meta-analysis is a poor reflection of the distribution of sampling errors of absolute values. The distribution of sampling errors will be highly skewed for modest estimates with substantial uncertainty (i.e., when $\sigma(m)_i$ is large relative to $|\hat{x}_i|$), while the mixed-effects meta-analysis assumes normal errors.

The 'analyse-then-transform' approach, i.e., of modelling the raw meta-data, i.e., the $\hat{x}_i$ values rather than the derived $|\hat{x}_i|$ values, and then deriving the mean absolute value, was unbiased across the majority of the range of parameter values. To some extent, this can be interpreted as the analysis being a match to the data-generating mechanism. It is true that I simulated the data under the statistical model that the mixed-effect meta-analysis applies to values of $\hat{x}_i$ and their associated standard errors. However, this type of model might in fact often be a very reasonable approximation to how values in many meta-datasets are obtained. This meta-analytic approach was slightly upwardly biased at the very lowest values of the true variance of x. This is because I constrained the estimate of $\sigma(x)$ to be positive, and so at the smallest true values of $\sigma(x)$, the estimate must be at least a slight over-estimate (in general, it is hard to imagine an estimator of a variance that is constrained to be positive, that will not be upwardly biased for small true values). Since the absolute value depends positively on the variance, this generates slight upward bias at the smallest true values.

Here, I have only focused on meta-analysis of the mean, and of the mean absolute values. There are of course many other quantities that may be of interest in a meta-analysis. Most quantities that are derived from quantities in the literature, according to a non-linear function, will be biased in informal and 'transform-then-analyse' meta-analyses. In addition to the mean (but not the mean absolute value), quantities such as regressions should generally be unbiased, even if sampling error is not explicitly considered. For example, consider a meta-dataset with

295 estimates of birds' singing rates from different studies. Suppose that standard errors of singing
296 rates were not available. We have seen that the estimate of mean singing rate would not be
297 biased in a summary statistic-based informal meta-analysis. Similarly, we should not expect an
298 inference of the average regression of singing rate on a predictor variable, such as a measure of
299 forest cover, to be biased in informal meta-analyses. In contrast, quantities such as variances,
300 mean absolute values, or the mean absolute differences among treatments, all depend on the
301 dispersion of values among studies, and will therefore be biased in informal meta-analyses, and
302 will also be biased in 'transform-then-analyse' approaches to formal meta-analysis.

303 # 5 Re-analyses of informal meta-analyses

304 ## 5.1 The average magnitude of natural selection

305 Kingsolver et al. (2001) reported on an informal meta-analysis of selection gradients and dif-
306 ferentials (Endler, 1986; Lande, 1979; Lande and Arnold, 1983). One of their most important
307 findings is that non-trivial directional selection is common in nature. They report an average
308 magnitude of variance-standardised directional selection gradients of 0.23 (the full distribution
309 is depicted in figure 2a)[3]. As we have seen (equation 6), this finding potentially represents a
310 substantial over-estimate, due to sampling error. The average standard error of selection gradi-
311 ent estimates in the database is about 0.15. So, in the improbable but instructive hypothetical
312 scenario where there was no selection in any study (just statistical noise arising from finite
313 sample size), the estimated mean absolute value of selection gradients that would be inferred
314 in an informal meta-analysis would be on the order of

$$\sqrt{\frac{2}{\pi}} \cdot 0.15 = 0.12.$$

315 **Re-analysis**

316 I used a mixed model to decompose the observed variation in selection gradients into that
317 arising from statistical noise and that which may represent real variation. The model took the

[3]There is a small difference in the mean absolute value of directional selection gradients in the database as a whole (0.23), and in that subset of the database that has standard errors (about 0.19). It probably arises from studies with very small sample size being over-represented in the portion of the database without standard errors.

form

$$\hat{\beta}_i = \hat{\mu_\beta} + m_i + e_i. \quad (10)$$

$\hat{\beta}_i$ are estimated selection gradients, and μ is the model intercept, or the estimated mean selection gradient. m_i are measurement errors, which are of course unknown, although we know they are drawn from estimate-specific distributions approximately following $m_i \sim N(0, SE_i^2)$. e_i are residuals, and are assumed to follow $e_i \sim N(0, \hat{\sigma}^2(\beta))$, where $\hat{\sigma}^2(\beta)$ is estimated. I then derived an estimate of the mean absolute value of selection as the mean of a folded normal distribution (equation 7) defined by the mixed-models estimates of $\hat{\mu_\beta}$ and $\hat{\sigma}^2(\beta)$. To produce a comparable mixed model-based analysis that does not account for sampling error, I also fitted the model

$$\hat{\beta}_i = \hat{\mu_\beta} + e_i. \quad (11)$$

I fitted both models using MCMCGLMM (Hadfield, 2010), using default diffuse priors. I then derived the mean absolute value of selection gradients as the expectation of a folded normal distribution defined by the parameters estimated in the models defined by equations 10 and 11.

Accounting for statistical noise generates an estimate of the variance of selection gradients of 0.0156 (i.e., from the model in equation 10; this is the posterior mode of the parameter in the mixed model; this statistic is used for estimates throughout), with a 95% credible interval of 0.0121 - 0.0207. By contrast, the model in equation 11 yields a variance of estimated selection gradients of 0.0775 (95% CI: 0.0689 - 0.0890). The corresponding standard deviations are 0.12 (95% CI: 0.11 - 0.14) and 0.28 (as for the estimate from the raw data, see above, with 95% CI: 0.26-0.30).

The model-based estimate of the average magnitude of selection gradients obtained as the mean of a folded normal distribution is 0.10 (95% CI: 0.09 - 0.12). The corresponding estimate based on the estimated selection gradients without accounting for sampling error is 0.23 (95% CI: 0.21 - 0.24), which closely matches the estimate obtained by simply calculating the mean of the absolute values of all the estimated directional selection gradients in the database.

While the purpose of the present work is not necessarily to perform a comprehensive re-

344 analysis of any given study, the average strengths of selection for different strata of the King-
345 solver et al. (2001) dataset are clearly of interest. I therefore ran the basic mixed model
346 analyses, with and without accounting for sampling error, for several major subsets of the
347 database, continuing to focus on directional selection gradients. Because (a) analyses are (cor-
348 rectly) much less apparently powerful when accounting for sampling error, and (b) sample sizes
349 for some strata are small and further reduced by incomplete reporting of the standard errors
350 necessary for meta-analysis, I did not conduct every possible analysis. Rather I subsetted the
351 database taxonomically for vertebrates, invertebrates, and plants, by trait type for life history
352 and morphology, and by fitness component for fecundity, mating success, and survival.

353 The general pattern that the magnitude of selection is inflated in analyses that do not
354 account for statistical noise at the level of individual estimates is supported at every level within
355 the database that I considered (table 1). Selection for life history traits is weakest, but this
356 probably reflects the definition used for life history traits. Many of the traits represent timing
357 in the life cycle, rather than life history traits *sensu stricto*, i.e., as in variables defined by a life
358 table. The general previously-reported patterns hold for means of selection gradients, which
359 are not expected to be biased by sampling error. Selection is generally positive for morphology,
360 and positive selection often acts through mating success (this may be primarily driven by
361 selection for morphology). Statistical noise at the level of the meta-analysis is increased (see
362 credible intervals reported in table 1), relative to the magnitudes of the estimates, in the formal
363 model that accounts for sampling error at the level of the component studies. This does not
364 represent a decrease in statistical power, but rather an improvement in realism relative to the
365 over-optimism of analyses that do not account for statistical noise.

366 The normal approximation to the distribution of selection gradients assumed in the residual
367 structure of a model such as that in equation 10 may generally provide a pragmatic and robust
368 approach to investigating components of variation in any observed dataset. However, we may be
369 interested in other aspects of the distribution. For example, it is very reasonable to think that
370 the true distribution of selection gradients may have thicker tails than the normal distribution.
371 I therefore constructed a model that is analogous to that in equation 10, except that the
372 underlying variation in selection gradients is modelled with a three parameter t-distribution.

This model takes exactly the same form as equation 10, except that the normal distribution from which the e_i are drawn is replaced by the three parameter t-distribution with mean zero (because the model contains an intercept), and estimated variance and degrees of freedom.

The distribution of selection gradients from the t-distribution based model is depicted in figure 2b. Comparison to figure 2a shows the dramatic difference between the distribution of *estimated* selection gradients and the underlying distribution of selection gradients. The inset figure depicts the relationship between unit variance-standardised trait values and relative fitness that is implied by the average magnitude of estimated selection gradients, which is very strong selection (see arguments in Hereford et al. 2004); $|\beta| = 0.22$ corresponds to approximately a 2.5-fold change in fitness over a range from two standard deviations below to above the mean phenotype. Such a selection gradient clearly does occur in nature (figure 2b), but is far rarer than the original informal meta-analysis suggested. The mean absolute magnitude of directional selection gradients in the t-distribution model[4] is 0.090 (95% CI: 0.076 - 0.108).

Other inferences about the mean absolute value of selection

Knapczyk and Conner (2007) argued that the mean magnitude of selection gradients in Kingsolver et al.'s meta-analysis was not inflated by sampling error. Their analysis relied on subsampling from a restricted array of very large datasets. This is a potentially very useful approach, but it relies on an assumption that the relevant properties of the restricted array of datasets are the same as in the larger database. Close inspection reveals that this cannot be the case in this instance. The restricted array of estimates of β in Knapczyk and Conner (2007) contains some very large selection gradients including $\beta = 1.12$ for selection of flower number via seed production, and three gradients of the fifteen in the Knapczyk and Conner (2007) dataset have an absolute value above 0.5.

Inspection of the raw data from the Kingsolver et al. (2001) database (Kingsolver et al.'s figure 5, figure 2a here), reveals that such large selection gradients are very far from representative of the data as a whole. The selection gradients in Kingsolver et al. (2001) have larger sampling errors, overall, than those in the Knapczyk and Conner (2007) dataset, and this larger sampling

[4] obtained as $\int |x| d(x|\mu, \sigma^2, k)\, dx$, where $d(x|\mu, \sigma^2, k)$ is the density of the three parameter t-distribution with mean μ, variance σ^2 and degrees of freedom k.

error can only inflate the apparent frequency of very large selection gradient estimates. If such large (true) selection gradients were similarly frequent in the study systems from which the Kingsolver et al. dataset was constructed, then similarly large (or larger) estimated selection gradients would be similarly common, and they are not (Figure 2a). Furthermore, the few selection gradient estimates of similar magnitude in the meta-database come exclusively from studies with very small sample size (Kingsolver et al., 2001) - precisely those that would be expected to yield estimates of large magnitude due to sampling error alone.

Note that Knapczyk and Conner (2007) made no errors that cause their dataset to be non-representative; it is simply by inspection of the distribution of estimates in the Kingsolver et al. (2001) database that it is apparent that no true underlying distribution of selection gradients, observed with sampling error, can be compatible with the high frequency of very large estimates in the Knapczyk and Conner (2007) analysis. The similarity between the results of Conner et al.'s analyses and the distribution of selection gradient estimates in the Kingsolver et al. (2001) dataset is coincidental, and does not conflict with the inevitability that sampling error will (potentially greatly) inflate estimates of the magnitude of effects in informal meta-analyses.

Hereford et al. (2004) clearly described the statistical mechanism by which sampling error can inflate inferenes of the mean magnitude of selection. They applied a post-hoc correction for sampling error using reported standard errors, and investigated the effect on the inference of the mean absolute values of selection gradients. Their correction was not expected to completely alleviate the problem, and the degree to which it solved the problem was not clear. Their partially-corrected estimate of the mean absolute value of selection gradients was consequently intermediate to that given by the original informal meta-analysis, and the formal model-based analysis presented here.

Finally, Kingsolver et al. (2012) reported on an effort to apply a formal meta-analysis to an updated database of selection gradient estimates. They performed several analyses of a database originally presented in Kingsolver and Diamond (2011), which combined datasets from Kingsolver et al. (2001) and Siepielski et al. (2009). Their position on the effects of accounting for error is unclear. They specifically state, with respect to quantities such as the mean absolute value of selection gradients, both that their results are similar to previous

studies, and also that there are large effects of accounting for error (which previous studies did not do).

Kingsolver et al. (2012)'s inference of the mean absolute value of selection, accounting for sampling error, is much greater than their inference based on a naive analysis (which they refer to as 'uncorrected $|\beta|$'). This is a mathematical impossibility, or at least could only occur if the properties of selection gradient estimates that are reported with and without standard errors are vastly greater than seems plausible. It seems likely that some error occurred in those analyses. My own re-analysis of the combined dataset reveals a mean absolute value of estimated selection gradients (i.e., via informal meta-analysis) of about 0.21, both for the subsets of the data with and without reported standard errors. This contrasts sharply with the the 'uncorrected' value of about 0.05 reported in Kingsolver et al. (2012). I was able to closely replicate their estimate of the mean $|\beta|$ from formal mixed effects meta-analysis (the analyse-then-transform approach) of about 0.14.

It may initially seem that the inference of the mean absolute value of selection from the combined Kingsolver et al. (2001) and Siepielski et al. (2009) databases should be superior, as it is based on a larger sample size. However, the credible intervals of the mean $|\beta|$ from the Kingsolver et al. (2001) and combined datasets do not overlap (95% CIs of 0.09 - 0.12 and 0.14 - 0.17, respectively). Therefore there must be some underlying difference between the two databases. Specifically, in that portion of the estimates from the Siepielski et al. (2009) study, which are temporally-replicated studies, must have stronger selection on average. I suspect that people will be mostly inclined to invest long-term efforts in studies of traits that they already know to be under selection. If this is the case, then the studies contributing to the original Kingsolver et al. (2001) dataset might give the best impression of the average magnitude of selection across a wide range of trait types and scenarios.

5.2 The frequency and magnitude of sexually antagonistic selection

Cox and Calsbeek (2009) present an informal meta-analysis of sexually antagonistic selection. They report that 41% of pairs of selection coefficient estimates, obtained for each sex for homologous traits, are sexually antagonistic, i.e., take opposite signs in the sexes. The standard

457 deviations of male and female selection coefficients (gradients and differentials combined) are
458 0.37 and 0.34, and the correlation between them is 0.19. The coefficient estimates are plotted
459 in figure 3a. The coefficient estimates that have associated standard errors are plotted in figure
460 3b.

461 The mean standard errors of selection coefficients are 0.17 for males and 0.20 for females.
462 The sex-specific sampling errors are expected to be uncorrelated, i.e., due to statistical noise
463 alone, there are few conditions in which studies that overestimate the true value of a selection
464 coefficient in one sex are no more or less likely to overestimate the corresponding coefficient
465 in the other sex. I simulated a set of random numbers, with one number corresponding to
466 every selection coefficient in the meta database that had a reported standard error. These
467 random numbers all had expectations of zero, and variances determined by the square of the
468 standard error. The distribution of these samples reflects the instructive though implausible
469 scenario of the distribution of *estimated* sex-specific selection coefficients that would arise in the
470 hypothetical situation where no selection occurred in either sex in any study from the literature.
471 Thus, this scenario can give some insight into the influence of sampling error alone on inferences
472 of the frequency of sexually-antagonistic selection. The distribution of these hypothetical data
473 points is given in figure 3c; in this scenario, statistical noise causes approximately 50% of
474 estimates to appear to be sexually antagonistic. A key feature of the pattern in figure 3c is
475 that, no matter how many estimates are included in the informal meta-analysis, a substantial
476 impression of sexually-antagonistic selection will result, as a result of sampling error at the
477 level of the individual studies.

478 We can treat the problem more formally. Cox and Calsbeek (2009) used a measure of
479 sexually-antagonistic selection based on the absolute difference between paired male and female
480 selection coefficients

$$\hat{SA}_i = |\hat{S}_m - \hat{S}_f| \tag{12}$$

481 where $\hat{S}_m$ and $\hat{S}_f$ are estimated male and female variance-standardised selection coefficients
482 (either differentials or gradients). Cox and Calsbeek (2009) provide a discussion of how this
483 coefficient relates to different aspects of sexually-antagonistic selection. If we assume that the

484 true distribution of selection coefficients in males and females is bivariate normal, and that
485 sampling errors of male and female selection gradients are both normal and uncorrelated, we
486 can derive an expression for the bias in an informal meta-analysis of sexually-antagonistic
487 selection.

488 The variance of the distribution of differences in true selection coefficients in males and
489 females is

$$\sigma^2(S_m - S_f) = \sigma^2(S_m) + \sigma^2(S_f) - 2\sigma(S_m, S_f) \tag{13}$$

490 where $\sigma^2(S_m)$, and $\sigma^2(S_f)$ are the variances in true selection coefficients in males and females,
491 and $\sigma(S_m, S_f)$ is the covariance of true selection coefficients. The variance of the distribution
492 of differences in estimated selection coefficients in males and females is

$$\begin{aligned}\sigma^2(\hat{S}_m - \hat{S}_f) &= \sigma^2(\hat{S}_m) + \sigma^2(\hat{S}_f) - 2\sigma(\hat{S}_m, \hat{S}_f) \\ &= \sigma^2(S_m) + \sigma^2(m)_{S_m} + \sigma^2(S_f) + \sigma^2(m)_{S_f} - 2\sigma(S_m, S_f),\end{aligned} \tag{14}$$

493 where $\sigma^2(m)_{S_m}$ and $\sigma^2(m)_{S_f}$ are the sampling variances of male and female selection coefficients.

494 The mean absolute value of the difference between two independent draws from the same
495 normal distribution is

$$E[|x_i - x_j|] = \frac{2}{\sqrt{\pi}}\sigma(x) \tag{15}$$

496 (Nair 1936, eq. 35). The bias in an informal meta-analysis of SA can therefore be written
497 using equations 13, 14 and 15

$$\begin{aligned}&\frac{2}{\sqrt{\pi}}\sqrt{\sigma^2(S_m) + \sigma^2(m)_{S_m} + \sigma^2(S_f) + \sigma^2(m)_{S_f} - 2\sigma(S_m, S_f)} - \frac{2}{\sqrt{\pi}}\sqrt{\sigma^2(S_m) + \sigma^2(S_f) - 2\sigma(S_m, S_f)} \\ =&\frac{2}{\sqrt{\pi}}\left(\sqrt{\sigma^2(S_m) + \sigma^2(m)_{S_m} + \sigma^2(S_f) + \sigma^2(m)_{S_f} - 2\sigma(S_m, S_f)} - \sqrt{\sigma^2(S_m) + \sigma^2(S_f) - 2\sigma(S_m, S_f)}\right).\end{aligned} \tag{16}$$

498 The expression is inelegant, but we can see that the quantity in brackets will be positive any
499 time that $\sigma^2(m)_{S_m}$ and/or $\sigma^2(m)_{S_f}$ are positive, which in practice will always be the case.

Re-analysis

I constructed a bivariate-response mixed model to partition (co)variation in sex-specific pairs of selection coefficients into portions arising from sampling error, and reflecting the underlying biological pattern. The model took the form

$$\begin{bmatrix} S_{m,i} \\ S_{f,i} \end{bmatrix} = \begin{bmatrix} \mu_m \\ \mu_f \end{bmatrix} + \begin{bmatrix} m_{m,i} \\ m_{f,i} \end{bmatrix} + \begin{bmatrix} e_{m,i} \\ e_{f,i} \end{bmatrix} \tag{17}$$

where $S_{m,i}$ and $S_{f,i}$ are the male and female-specific estimates for pairs of selection coefficients[5] indexed by i. Sampling errors are assumed to be drawn according to

$$\begin{bmatrix} m_{m,i} \\ m_{f,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} SE_{m,i}^2 & 0 \\ 0 & SE_{f,i}^2 \end{bmatrix} \right)$$

and residuals according to

$$\begin{bmatrix} e_{m,i} \\ e_{f,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2(m) & \sigma(m,f) \\ \sigma(m,f) & \sigma^2(f) \end{bmatrix} \right)$$

where residual variances and covariance of male and female selection gradients, $\sigma^2(m)$, $\sigma^2(f)$, and $\sigma(m,f)$, as well as the sex-specific means in equation 17 are estimated parameters. I implemented the model in JAGS (Plummer, 2010), with diffuse normal priors on the sex-specific means and a redundant prior parameterisation on the residual covariance matrix of selection coefficients.

The mean selection coefficient in each sex is positive: males: 0.092 (95% CI: 0.040 - 0.153), females: 0.074 (95% CI: 0.030 - 0.108). Critically, male and female selection coefficients covary strongly and positively. The residual covariance matrix obtained by fitting the model described

[5] The analysis is conducted on a mix of selection differentials and gradients, following Cox and Calsbeek 2009. This combination is reasonable as the values are all variance standardised.

in equation 17 (95% CIs in brackets) is

$$\begin{bmatrix} \sigma^2(m) = 0.067\ (0.054 - 0.106) & \sigma(m,f) = 0.038\ (0.024 - 0.054) \\ r(m,f) = 0.794\ (0.666 - 0.928) & \sigma^2(f) = 0.029\ (0.016 - 0.045) \end{bmatrix};$$

note that the sub-diagonal element is reported as the correlation. The consequence of this positive correlation of male and female coefficients is that sexually antagonistic selection is rare, and when it occurs, it is typically not highly antagonistic. Simulated values drawn form the inferred joint distribution of male and female selection coefficients are plotted in figure 3d. The proportion of pairs of selection coefficient estimates that differ in sign[6] is 20% (95% CI: 12 - 25%). Furthermore, when selection is sexually antagonistic, it is also weakest.

Figure 4 shows the distributions of two possible metrics of sexually-antagonistic selection. These metrics are calculated both from the raw data, i.e., by informal meta-analysis, and calculated from the 'analyse-then-transform' analyses made possible by the bivariate response random regression model. The first metric (figure 4a) is the distribution of products of male and female selection coefficients. This quantity is negative when selection takes different signs in the two sexes, and positive when selection is of the same sign. Values near zero indicate that there is little selection in one or both sexes. The second metric (figure 4a) is Cox and Calsbeek (2009)'s measure based on the absolute value of differences in male and female coefficients.

The model specified by equation 17 does not account for different levels of non-independence in the data. Accounting for statistical non-independence is not expected (on average, i.e., the analysis presented to this point is not expected to be biased) to change the inference about the underlying variance and covariance of sex-specific selection coefficients. However, accounting for non-independence may change our impression of how precisely we have characterised any given overall effect. A source of non-independence considered by Cox and Calsbeek (2009) is that pairwise reports of sex-specific selection coefficients from the same study tend to be

[6] obtained by $\int\int \frac{S_m \cdot S_f}{|S_m| \cdot |S_f|} \cdot N([S_m, S_f]^T, \mu, \sigma) dS_m dS_f$, where μ and σ are the mean vector and covariance matrix of selection coefficients.

similar. I therefore fitted the model

$$\begin{bmatrix} S_{m,ij} \\ S_{f,ij} \end{bmatrix} = \begin{bmatrix} \mu_m \\ \mu_f \end{bmatrix} + \begin{bmatrix} m_{m,ij} \\ m_{f,ij} \end{bmatrix} + \begin{bmatrix} r_{m,ij} \\ r_{f,ij} \end{bmatrix} + \begin{bmatrix} e_{m,ij} \\ e_{f,ij} \end{bmatrix} \tag{18}$$

where r denotes study, and j indexes the studies to which individual records belong. As above, the upper left elements are variances associated with male selection coefficients, the bottom right correspond to female selection coefficients, and the entries above the diagonal are covariances, and below the diagonal are correlations. The covariance matrix from which the r values are assumed to come is constructed and estimated equivalently to the residual covariance matrix (described above), and all other model components are treated as they were for the model described by equation 17.

The between-study and within-study covariance matrices of paired sex-specific selection coefficients are

$$\begin{bmatrix} 0.034\ (0.009-0.066) & 0.021\ (0.004-0.055) \\ 0.996\ (0.504-1.000) & 0.025\ (0.009-0.069 \end{bmatrix}, \text{ and } \begin{bmatrix} 0.041\ (0.029-0.071) & 0.015\ (0.007-0.028) \\ 0.678\ (0.398-0.901) & 0.012\ (0.005-0.022) \end{bmatrix},$$

respectively. The male variance is in the top left and the female variance is in the bottom right. 95% CIs are in brackets. The sub-diagonal element are the correlations. The total (co)variances and correlations are thus

$$\begin{bmatrix} 0.075\ (0.049-0.119) & 0.045\ (0.021-0.075) \\ 0.755\ (0.496-0.905) & 0.043\ (0.021-0.084) \end{bmatrix}.$$

Accounting for non-independence among data points that come from the same studies therefore does not appreciably change the overall pattern. The credible intervals of the total variance components obtained from the second model are slightly larger and are probably more appropriate. Differences in whether or not selection is sexually antagonistic or not seem to arise more from differences among traits, than from differences among studies.

Sexual dimorphism and sexually antagonistic selection

Cox and Calsbeek (2009) considered whether any association exists between sexual dimorphism and sexually antagonistic selection. This is a very interesting problem. A negative relationship

between these phenomena might indicate that the evolution of sexual dimorphism generally has resolved sexual conflict, while a positive relationship would indicate a general pattern of ongoing conflict between the sexes. In the context of the analyses pursued to this point, a relationship between sexual dimorphism and sexually antagonistic selection would primarily be manifested as a (statistical) dependence between sexual dimorphism and the covariance between male and female selection coefficients. Methods for estimating the dependence of a covariance on a continuous variable are not well developed.

Standard modelling procedures do not exist to accommodate hypotheses about how covariance structures vary according to continuous variables. Therefore, determining how typical magnitudes of sexually antagonistic selection covaries with a predictor such degree of sexual dimorphism would deserve an independent study in itself. Here I make only a preliminary attempt. A model structure that may be pragmatic would be to treat the correlation of male and female selection gradients as a continuous function of the degree of sexual dimorphism, and model the shape of that function as a sigmoidal relationship ranging between -1 and +1. I therefore parameterised the correlation as

$$r_{S_m,S_f,i} = \frac{2e^{a+b \cdot D_i}}{1 + e^{a+b \cdot D_i}} - 1 \tag{19}$$

where α and b are the regression parameters controlling the shape of the logistic curve that is scaled between negative and positive one (note that $\frac{e^{a+b \cdot D_i}}{1+e^{a+b \cdot D_i}}$ would represent a logistic curve between 0 and 1). $r_{S_m,S_f,i}$ can then be thought of as the correlation that would be observed among a group of paired sex-specific selection coefficients, all from systems with sexual dimorphism D_i. I used the absolute value of the measure of sexual dimorphism available in the Cox and Calsbeek (2009) database, which is the difference between sex-specific means. I specified the variances of the sex-specific selection coefficients independently, and then obtained the dimorphism-dependent covariance of paired sex-specific selection coefficients as $r_{S_m,S_f,i}\sqrt{\sigma^2(m)}\sqrt{\sigma^2(f)}$.

The parameters of the regression of $r_{S_m,S_f,i}$ on the degree of sexual dimorphism are α: 2.2 (95% CI: 0.5 - 4.3), and b: 2.1 (95% CI: -7.8 - 25.0). About 80% of the posterior distribution

584 of b is greater than zero. Thus the overall pattern appears to be for sexual dimorphism to be
585 associated with a reduction in the degree of sexually antagonistic selection, although the value
586 of the coefficient controlling this pattern has a posterior distribution that substantially overlaps
587 zero. It is not surprising that this regression has a very large standard error. Considering that
588 each pair of estimates does not provide a concrete datapoint, but rather a very uncertain
589 inference about sexually-antagonistic selection, the formal meta-analysis may correctly have
590 great uncertainly in measures that seem easily estimable in an informal meta-analysis. The
591 correlation between male and female coefficients in the absence of sexual dimorphism is thus
592 about 0.85, while at higher levels of dimorphism, the correlation approaches one.

593 **5.3 Population and species differences in reaction norm shape**

594 Murren et al. (2014) report on differences between average values, slopes, curvatures, and
595 higher-order aspects of the shapes of reaction norms between species and populations. Their
596 primary conclusions include (1) that shapes, i.e., slopes and curvatures, of reaction norms
597 evolve more than average trait values, and (2) that curvature of reaction norms evolves more
598 than the slope. Statistical noise will inflate apparent differences between parameters such as
599 means[7], slopes and intercepts. Furthermore, depending on the scaling of the environmental
600 variables, statistical noise will contribute differently to apparent variation in means, slopes and
601 curvatures. Therefore, sampling error alone will create specific patterns in the mean absolute
602 differences of averages, slopes, and curvatures of pairs of reaction norms.

603 A simple simulation may be instructive. Again, we will start with a simple situation with
604 trivial biology, and focus on how unbiased statistical noise in the literature may be converted
605 into superficially, and misleadingly, biologically interesting patterns in a naive meta-analysis.
606 Assume that some large number of studies are conducted, and that in each, two populations
607 are assayed for mean phenotype in each of three (ordered) environments. Assume that every

[7]Here, four different words will be used for aspects of the average value of a reaction norm. The mean will represent the population mean, which is the mean value of the reaction norm weighted by the distribution of the environment that the population experiences. The offset will refer to the mean value, weighting all values (given some range) of the environment equally. The intercept will be the value of the reaction norm at a given value of the environmental variable that is defined as the origin. The intercept is the same as the mean when the environmental variable is symmetrically distributed about the origin, and the reaction norm is linear. The intercept is the same as the offset when the environmental variable is centred on the origin, and the reaction norm is linear. The means and offsets can be calculated for non-linear reaction norms, and this will be done as appropriate. The term 'average' will be used to refer to these values collectively, when the distinctions are not critical.

608 population in every study and in every environment has the same mean value (the mean value
609 is actually irrelevant), and that the standard error of the mean is 1 unit in every case (this
610 value is also completely irrelevant to the pattern that results, so long as it is non-zero). For
611 this null scenario, I simulated data, and calculated the difference in means between populations
612 (species) for each of the simulated studies, as well as the differences in slopes and curvatures,
613 following the expressions used by Murren et al. (2014). The distribution of the magnitudes,
614 i.e., absolute values, of these differences is plotted in figure 5. Murren et al. (2014) report
615 estimates of mean absolute differences in reaction norm components from an analysis that is
616 weighted by (the square root of) sample size. Note that weighting does not solve the problem
617 illustrated here. A well-designed weighting scheme will be analogous to the transform-then-
618 analyse approach to meta-analysis, which can perform poorly for arbitrary derived quantities
619 (figure 1). Consider that these simulations assume equal error across all estimates, which may
620 occur if (among other things) there are equal sample sizes. As such, weighting by sample size
621 would provide a trivially identical result to an unweighted analysis, and the spurious pattern
622 would remain.

623 The pattern in figure 5 can also be obtained analytically. Again, I will focus on the scenario
624 where there are three environmental treatments, as these dominate the available data. Assume,
625 as above, that a pair of reaction norms (e.g., a congeneric or conspecific pair) are identical. Let
626 the mean phenotypes in the three environments for one population be denoted $\hat{x}_1$, $\hat{x}_2$, and $\hat{x}_3$,
627 and denote the corresponding three estimated mean phenotypes in the other population with
628 $\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_3$. Assume that all mean values are estimated with the same precision, such that
629 $\hat{x}_i \sim N(\mu, \sigma(m))$, $\hat{y}_i \sim N(\mu, \sigma(m))$.

630 The variance of the mean of the $\hat{x}$ or $\hat{y}$ values is

$$\sigma^2(\bar{\hat{x}}) = \sigma^2(\bar{\hat{y}}) = 3\left(\frac{1}{3}\right)^2 \sigma^2(m) = \frac{1}{3}\sigma^2(m) \tag{20}$$

631 which is simply the variance of three independent random values, each with the same variance.
632 The average of the slopes of the two line segments in each reaction norm is $\frac{1}{2}(\hat{x}_2 - \hat{x}_1) + \frac{1}{2}(\hat{x}_3 -$
633 $\hat{x}_2) = \frac{1}{2}\hat{x}_3 - \frac{1}{2}\hat{x}_1$ (or equivalent expressions with $\hat{y}$). Therefore the sampling variance of average

slopes is

$$\sigma^2(\hat{x}_i - \hat{x}_{i-1}) = 2\left(\frac{1}{2}\right)^2 \sigma^2(m) = \frac{1}{2}\sigma^2(m). \quad (21)$$

Curvature (defined by Murren et al. 2014 as the difference of slopes between adjacent intervals) for a study with three points is

$$(\hat{x}_3 - \hat{x}_2) - (\hat{x}_2 - \hat{x}_1) = \hat{x}_3 - 2\hat{x}_2 + \hat{x}_1$$

and so the variance in curvatures is

$$\sigma^2((\hat{x}_{i+1} - \hat{x}_i) - (\hat{x}_i - \hat{x}_{i-1})) = 2\sigma^2(m) + 2^2\sigma^2(m) = 6\sigma^2(m). \quad (22)$$

The mean difference between different reaction norm components is given by the expression $\frac{2}{\sqrt{\pi}}\sigma$, just as we used for the mean difference in male and female selection coefficients. Consequently, in the absence of any differences in reaction norms between conspecific or congeneric populations, a pattern in estimated mean differences in means, slopes, and curvatures will arise by sampling error alone. In our toy model, the pattern will be:

$$\frac{2}{\sqrt{\pi}}\sqrt{\frac{1}{3}\sigma^2(m)}$$

for means

$$\frac{2}{\sqrt{\pi}}\sqrt{\frac{1}{2}\sigma^2(m)}$$

for slopes, and

$$\frac{2}{\sqrt{\pi}}\sqrt{6\sigma^2(m)}$$

for curvatures. This pattern will be super-imposed on any true differences among these properties of reaction norms in nature.

Re-analysis

Distributions of intercepts, slopes, and curvatures can be modelled using mixed effects models, just as differences in mean values can, and were, in the preceding examples. To obtain model-

based estimates of differences in properties of reaction norms, I fitted the model

$$\begin{aligned} x_{ijk} =& A + B \cdot E_j + C \cdot E_j^2 \\ &+ a_{r,k} + b_{r,k} \cdot E_i + c_{r,k} \cdot E_i^2 \\ &+ a_{s,j} + b_{s,j} \cdot E_i + c_{s,j} \cdot E_i^2 \\ &+ a_{p,i} + b_{p,i} \cdot E_i + c_{p,i} \cdot E_i^2 \\ &+ e_i. \end{aligned} \tag{23}$$

This is a quadratic *random regression mixed model*. x_{ijk} are the environment-specific estimated mean values, and E_i are the corresponding values of the environmental covariate (expressed as treatment intervals in the raw data). I standardised the environment-specific estimated means in two ways. Murren et al. (2014) divided by the overall mean, and I did this as well. Furthermore (and see discussion below) a scaling that may better facilitate inference of both intraspecific and congeneric variation in reaction norms is to log (actually $ln(x+1)$, as there are zero values in the data) transform, and so I used logged data as well. i indexes studies, and j indexes paired estimates within studies. A, B, and C are the average intercept, slope, and curvature. The a, b, and c terms are the study-specific (or rather trait within study) random intercept, slope and curvature terms, associated with study r, species s, and population p. I modelled these terms as being drawn from the multivariate normal distribution

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix}_{x,y} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2(a) & \sigma(a,b) & \sigma(a,c) \\ \sigma(a,b) & \sigma^2(b) & \sigma(b,c) \\ \sigma(a,c) & \sigma(b,c) & \sigma^2(c) \end{bmatrix}_y \right)$$

where the parameters of the covariance matrix of a_i, b_i, and c_i values are estimated parameters, with $x \in \{k, j, i\}$ and $y \in \{r, s, p\}$. I modelled the residuals as coming from a common distribution, i.e., $e_{ij} \sim N\left(0, \sigma^2(e)\right)$.

I have preferred Bayesian approaches for all analyses (except simulations) to this point. While the random regression mixed model of variation in reaction norms can be fitted in a Bayesian analysis, I found that its results were extremely sensitive to prior specifications for the

668 variance components. This is not surprising (with hindsight), because only studies with four or
669 more environmental treatments can contribute to inferences about intercepts, slopes curvatures,
670 and residual variance. To avoid the need to use essentially arbitrary priors, I fitted this model
671 by restricted maximum likelihood, using `lme4` (Bates et al., 2014). Standard errors for variance
672 components in random regression models are not easily obtained from this software, and in
673 any case can be misleading when variance components are small and imprecisely estimated. I
674 therefore report only the (restricted) maximum likelihood estimates of the parameters of the
675 simplest model that reports parameters that are analogous to the main quantities reported by
676 Murren et al. (2014). These should be interpreted in the light that, given the model and the
677 currently-available data, the inferences about curvature are highly uncertain.

678 The scaling of the environmental variable, E in equation 23, is important to consider. Mur-
679 ren et al. (2014)'s calculations of means, slopes, and curvatures assume that all intervals be-
680 tween environmental treatments have equal meaning. This is one of two potential treatments.
681 Assuming equal biological meaning of all intervals assumes that those studies that use fewer
682 environmental treatments cover a proportionately smaller portion of the relevant range of the
683 environmental variable. I think that an alternative treatment may be more sensible. It seems to
684 me more likely that, on average, most studies are designed to span most of the relevant range of
685 environmental conditions, whatever that range may be for the study, species, populations, en-
686 vironmental variable, and traits in question. If this second option represents a more reasonable
687 model of how reaction norm studies are generally designed, the consequences of assuming equal
688 scaling of intervals, rather than equal scaling of the total environmental range, may be serious.
689 If two studies covered the same range of the environment, the one with fewer increments of
690 environmental conditions within that range would have greater calculated slopes and curva-
691 tures than the study with more increments, and thus would also have relatively exaggerated
692 differences between slopes and curvatures if equal scaling of increments was assumed.

693 Because neither treatment of the environmental variables is an obviously superior approach
694 for every study in the database, I applied both standardisations. These can be seen as useful
695 extremes, with truths for how each study was designed typically lying somewhere in between.
696 First (my *a priori* preference), I standardised the environmental variable in each study to

span the range from -2 to +2. The exact bounds are not necessarily important, although I chose -2 and +2 on the grounds that it might very roughly put the environmental variable in units of standard deviations, under the supposition most researchers will design studies with environmental variation that span the approximate limits of meaningful variation. If 'meaningful variation' is approximately normally distributed, 2 SD units spans most of the range. As a second treatment, that reflects Murren et al. (2014)'s assumptions, I mean-centred the environmental covariates, giving each increment equal value of one unit.

The model described by equation 23 does not explicitly account for sampling error. Rather, the different major potentially biasing factors (statistical noise, variation among treatments not associated with the focal reaction norm, and variation over and above quadratic effects) are treated together by the residual variance, in this case. The residual variance therefore combines these three major effects. The core difference between the quadratic random regression model, and the Murren et al. (2014) analysis is that there is some place, other than complexity in the form of reaction norms, for variation over and above that associated with reaction norms to be represented. It would be preferable to specifically model statistical noise; as it is, there will still be some effect of statistical noise to inflate inferences of reaction norm shape evolution. However, the standard errors necessary to explicitly model statistical noise are inconsistently reported in the literature, and as the relative amount of error in mean phenotype estimates is typically substantially smaller than that which occurs in selection coefficient estimates (see examples above), the effect could be modest. The analyses that I present should thus be considered conservative relative to my assertion that reaction norm shape evolution should be much more modest than reported by Murren et al. (2014).

The most immediately relevant variance components of the fitted mixed model defined by equation 23 are given in table 2. These model parameters represent variation among reaction norms. Mean absolute differences in intercepts, slopes, and curvatures are monotonic functions of the variance (true variance and/or sampling variance) according to $E[|x_i - x_j|] = \frac{2}{\sqrt{\pi}}\sigma(x)$ (see above). As such, the variances of intercepts, slopes and curvatures are the first pieces of information that the random regression mixed models provide about the relative importances of evolution of intercepts, slopes and curvatures. Under both standardisations, variation in

intercepts is the major component of variation in intercepts, both among species (table 2a) and among populations (table 2b). Transformation of these variances can put the relationships in a slightly different terms, that might also be useful for interpretation, and that relate more directly to the quantities (mean absolute differences) reported by Murren et al. (2014). In table 2d,e the mixed model results are reported in terms of mean absolute differences, and the results for curvature are reported as mean absolute differences in second derivatives. There is no overall pattern for reaction norm evolution to be dominated by evolution of reaction norm shape, although evolution of reaction norm shape among species may be somewhat more important than among populations. All these interpretations should be made keeping in mind that a modest quantity of data contributes to the inferences about variation in reaction norm curvatures.

The variances of reaction norm parameters among congeneric species, as estimated form the mixed model, has a different interpretation than the quantity estimated with summary statistics by Murren et al. (2014). Because any data from a given species necessarily is collected on individuals from some population within that species, the summary statistic-based approach includes both among-population and among-species variation in the inferences about congeneric differences in reaction norms. In contrast, the species-level variation inferred from the mixed model analysis is more hierarchical, representing the variation attributable to species.

Probably the best way to visualise the information about evolution of quadratic reaction norms that is contained in the fitted mixed models is by predictive simulation. Figures 6 and 7 show simulated pairs of reaction norms (with environmental variables standardised to common ranges), for intra-specific and congeneric reaction norms, respectively. Thus, these are not fitted results for any specific pairs of reaction norms in the meta-dataset, but rather, these are visualisations of the fitted model, converted for presentation into a format that closely corresponds to the main biological questions. Figures 6 and 7 show that among-species differences in reaction norm shapes are indeed generally greater than within-species differences. While reaction norms do vary in shape at both levels, most differences are in the mean, especially in the centre of the ranges of the reaction norms, where the quadratic form of the random regressions should provide the most reasonable approximations.

6 Discussion

The primary goal of this article is to highlight the conditions under which it is necessary to account for the observation process in synthetic meta-analysis, and how this can be accomplished with mixed models. In support of this goal, I suggest that many quantities of potential meta-analytic interest might best be obtained by modelling the distribution of quantities that are reported in the literature (rather than quantities derived from literature reports), and subsequently using these models to address biological questions. It should be clear that many meta-analytic questions, especially those relating to average magnitudes (or average magnitudes of differences, as in the second and third example re-analyses) absolutely require procedures that can separate biological signal from statistical noise. It must be stressed that, in each of the three examples, the results presented here and their modified interpretation are not a result of more powerful analyses. Even with infinite sample size (i.e., number of studies in a meta-dataset) the misleading conclusions of the informal meta-analyses would have occurred.

Importantly, it has been possible to clarify that there are conditions under which meta-analyses that do not account for statistical error will be biased. Meta-analytic quantities that do not depend on the dispersion of the values reported in the literature should generally fall into this category. This may be a useful finding in itself. Quantitative information about uncertainty, e.g., standard errors, are not universally reported, and in fact are disappointingly inconsistently reported in some literatures (e.g., in analyses of natural selection). While meta-analytic inferences of a given dataset will always be more precise if differences in precision among studies are taken into account, formal meta-analyses may not necessarily be most powerful when a choice must be made between a large dataset without, and a smaller dataset with, standard errors.

In the course of developing the mixed model-based meta-re-analyses, several useful biological results have come to light. First, the average magnitude of selection gradients is likely not as large as has been reported. In fact, the average magnitude of selection gradients as estimated in the analyse-then-transform meta-analysis is approximately half (0.10 vs. 0.19 or 0.23, depending on what subset of the data is considered) that which was previously reported. This is a

783 rather substantial difference in terms of interpretations of potential rates of adaptive evolution,
784 and a very substantial difference in terms of the size of studies that may need to be designed to
785 characterise typical strengths of selection in the wild. Second, the frequency at which sexually
786 antagonistic selection occurs is probably much less than that suggested by summary statistics
787 of paired estimated sex-specific selection coefficients. Furthermore, when sexually antagonistic
788 selection does occur, it is far more subtle than the impression given from considering the joint
789 distribution of male and female selection coefficient estimates. Third, evolution of reaction
790 norms is not generally dominated by evolution of their shape. In fact the formal meta-analysis
791 yields the opposite qualitative finding to that of the informal analysis: at least at the popula-
792 tion level, most trait evolution seems to be of mean values across environments, particularly
793 for divergence among conspecific populations.

794 None of these new findings should be viewed as a negative result. Relatively more mod-
795 est selection than is suggested by summary statistics applied to *estimated* selection gradients
796 goes some way toward explaining stasis (Merilä et al., 2001; Walsh and Blows, 2009), at least
797 in general terms. In practical terms, the approximate halving of the inference of the typical
798 strength of selection means that the sample sizes required to characterise 'typical' selection will
799 be quadrupled, following power calculations such as those in Hersch and Phillips (2004). Simi-
800 larly, it is useful to know that patterns of sexual antagonism (note that, in general, homologous
801 traits generally have very high genetic correlations between the sexes; Poissant et al. 2010) may
802 generally be much more subtle than is suggested by the main high profile results on the topic
803 (for e.g., Chippindale et al. 2001 and Foerster et al. 2007). Finally, the revised finding that
804 reaction norm shapes are not incredibly evolutionarily labile may be an interesting indication
805 that developmental systems are relatively stable (see also Voje et al. 2014).

806 Some statistical procedures may seem initially useful for dealing with sampling error in
807 meta-analysis. First, it is important to note that the issues discussed here are not a result
808 of a lack of statistical hypothesis testing in previous meta-analyses. Only formal statistical
809 methods that account for observation processes, as necessary for the specific goals of a given
810 meta-analysis, will prevent white noise at the level of individual datasets from being converted
811 into severe biases in meta-analyses. Second, weighting by sample size, the inverse of standard

errors, or other aspects of precision, will not necessarily solve the problems discussed here, when the interest in a meta-analysis is in any feature other than the mean of a phenomenon. Formal meta-analytic weighting methods, e.g., the method of moments estimators of means and variances (reviewed in Rosenberg 2013) will perform very similarly to the transform-then-analyse mixed model approach in the simulation section of this paper (dotted line in figure 1) when applied to derived quantities that depend on the dispersion. Third, subsetting meta-data to consider only statistically significant results may seem like a way to make inferences using only the most reliable portion of a meta-dataset, but such a practice will generally make the problems much worse. The subset of results in any literature that are statistically significant will generally provide very upwardly biased impressions of the magnitudes of phenomena (Gelman and Weakliem, 2009).

How is one to know if some specific inference will be biased by statistical noise in a meta-analysis? For each of the three examples I re-analysed, instructive analytical results about bias was obtainable (typically for simplified, but instructive, models). However, for other meta-analyses of the many potentially complex but interesting quantities that may be of interest in ecology and evolution, analyses such as these may not be tractable. Two useful guiding principles should be that: (1) biases should arise if the quantity of interest in an aspect of the dispersion (e.g., standard deviation, variance, mean difference) of quantities that are reported in the literature (see for e.g., Morrissey and Hadfield 2012), and (2) if the quantity of interest is obtained from a non-linear transformation (e.g., absolute value) of the quantities that are reported in the literature. A simulation approach may be useful in any specific situation. Before or after a meta-dataset is assembled, one can simulate some biologically null (or otherwise) "true" values, and then generate simulated estimates by adding error to those simulated true values (these errors can be drawn from distributions defined by standard errors, if available). Researchers can then apply their meta-analytic methods (informal or otherwise) to these simulated data to check whether sampling error causes appreciable deviation from their simulated patterns. This is the procedure that I did in the simple simulations to demonstrate how sampling error would affect the informal meta-analyses of sexually-antagonsitic selection (figure 3) and variation in reaction norms (figure 5). This type of simulation led to the deletion

of a meta-analysis of measures of spatial autocorrelation (e.g., of Moran's I, which is a complex transformation raw data from each study) in selection from Siepielski et al. (2013), as it uncovered severe biases arising from sampling error and non-random selection of study sites.

Further developments of meta-analytic techniques may be required for analysis of many parameters of interest in evolutionary biology. In this paper, I have focused on analysis of quantities that are non-linear transformations of quantities that are reported in the literature. Another class meta-analytic problems that is worthy of more methodological attention may be the analysis of bounded quantities. For example, meta-analysis of variance may potentially be of interest, but variances cannot (typically) be less than zero. Consequently, sampling errors of variance estimates will be asymmetric, potentially causing bias (similarly to simulations herein for the transform-then-analyse approach; figure 1). For variances, Nakagawa et al. (2015) have suggested that analyses could be conducted on the log scale. Results of such log-scale analyses could subsequently be transformed back into the original scale, if desired. Another situation where conducting meta-analyses on a different scale (and subsequently transforming results) could prove useful is in analysis of quantities such as heritability (e.g., see informal meta-analyses in Postma 2014) and other estimates of phenomena that are biologically useful to express as bounded quantities (e.g., measures of reproductive isolation, Sobel and Chen 2014, or phenotypic or genetic correlations). Means for transformation of estimates and their sampling variances to a scale where errors will be symmetric are not currently obvious in such cases.

Additional development of the "analyse-then-transform" approach to meta-analysis advocated here may be very useful as well. For meta-analytic inferences such as those made here, derived quantities (e.g., the mean magnitude of selection) may depend on complexities of the distribution of untransformed quantities. It is reassuring that the analyses assuming normal distributions and t-distributions of directional selection gradients yielded very similar inferences of the average magnitude of selection. It seems plausible that inferences based on normal distributions might typically be quite pragmatic. However, it should not be surprising if situations arise where the use of much more flexible random distributions in meta-analysis (Higgins et al., 2009) proves useful or even necessary.

The surge in popularity of meta-analysis may be occurring at the cost of qualitative synthesis. There is probably a great deal that can be gained from considering the expert opinion of a person who has invested time and thought in a particular topic. Much of what can be gained by qualitative review may easily be missed in the developing paradigm where synthesis is achieved primarily via meta-analysis. The insight provided by those rare studies that are particularly cleverly designed so as to strike at the core of an outstanding issue is greatly diluted in a meta-analysis. The most creative and insightful studies may even be excluded from meta-analyses, if they rely on particularly clever, but non-standard, approaches. We should not dismiss the service provided to any given field by a dedicated worker determining just what specific qualitative insights may be buried in large literatures.

Acknowledgements

Sam Scheiner, Thomas Hansen, Graeme Ruxton, Maria João Janeiro, Andy Gardner, Mike Ritchie, Jarrod Hadfield, Julia Koricheva, Kerry Johnson, and Shinichi Nakagawa provided insightful comments, useful discussions, and patient tolerance of my neuroticisms, on this topic. I am supported by a University Research Fellowship from the Royal Society (London).

References

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

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1-7 ed.

Chippindale, A. K., J. R. Gibson, and W. F. Rice. 2001. Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in *Drosophila*. Proceedings of the National Academy of Sciences 98:1671–1675.

Cox, R. M., and R. Calsbeek. 2009. Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. The American Naturalist 173:176–187.

895 Endler, J. A. 1986. Natural selection in the wild. Princeton University Press.

896 Foerster, K., T. Coulson, B. C. Sheldon, J. Pemberton, T. H. Clutton-Brock, and L. E. B.
897 Kruuk. 2007. Sexually antagonistic genetic variation for fitness in red deer. Nature 447:1107–
898 1111.

899 Gelman, A., and D. Weakliem. 2009. Of beauty, sex and power. The American Scientist
900 97:310–316.

901 Glass, G. V. 1976. Primary, secondary and meta-analysis of research. Educational Researcher
902 5:3–8.

903 Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analysis. Ecology
904 80:1142–1149.

905 Hadfield, J. 2010. MCMC methods for multi-response generalized linear mixed models: The
906 MCMCglmm R package. Journal of Statistical Software 33:1–22.

907 Hereford, J., T. F. Hansen, and D. Houle. 2004. Comparing strengths of directional selection:
908 how strong is strong? Evolution 58:2133–2143.

909 Hersch, E. I., and P. C. Phillips. 2004. Power and potential bias in field studies of natural
910 selection. Evolution 58:479–485.

911 Higgins, J. P. T., S. G. Thompson, and D. J. Spiegelhalter. 2009. A re-evaluation of random-
912 effects meta-analysis. Journal of the Royal Statistical Society 172:137–159.

913 Kingsolver, J. G., and S. E. Diamond. 2011. Phenotypic selection in natural populations: What
914 limits directional selection? The American Naturalist 177:346–357.

915 Kingsolver, J. G., S. E. Diamond, A. M. Siepielski, and S. M. Carlson. 2012. Synthetic analyses
916 of phenotypic selection in natural popualtions: lessons, limitations and future directions.
917 Evolutionary Ecology in review.

918 Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, C. Vignieri, D. Berrigan, E. Hill, A. Hoang,
919 P. Gilbert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations.
920 The American Naturalist 157:245–261.

921 Knapczyk, F. N., and J. K. Conner. 2007. Estimates of the average strength of natural selection
922 are not inflated by sampling error or publication bias. The American Naturalist 170:501–508.

923 Koricheva, J., and J. Gurevitch. 2013*a*. Place of meta-analysis among other methods of research
924 synthesis, in Handbook of Meta-analysis in Ecology and Evolution, J. Koricheva and J.
925 Gurevitch and K. Mengersen, eds., chap. 1. Princeton University Press, Princeton, New
926 Jersey.

927 ———. 2013*b*. Using other metrics of effect size, in Handbook of Meta-analysis in Ecology
928 and Evolution, J. Koricheva and J. Gurevitch and K. Mengersen, eds., chap. 7. Princeton
929 University Press, Princeton, New Jersey.

930 Koricheva, J., J. Gurevitch, and K. Mengersen. 2013. Handbook of meta-analysis in ecology
931 and evolution. Princeton University Press, Princeton, New Jersey.

932 Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body
933 size allometry. Evolution 33:402–416.

934 Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters.
935 Evolution 37:1210–1226.

936 Merilä, J., B. C. Sheldon, and L. E. B. Kruuk. 2001. Explaining stasis: microevolutionary
937 studies in natural populations. Genetica 122-222:112–113.

938 Morrissey, M. B., and J. D. Hadfield. 2012. Directional selection in temporally replicated studies
939 is remarkably constant. Evolution 66:435–442.

940 Murren, C. J., H. J. Maclean, S. E. Diamond, U. K. Steiner, M. A. Heskel, C. A. Handelsman,
941 C. K. Ghalambor, J. R. Auld, H. S. Callahan, D. W. Pfennig, R. A. Relyea, C. D. Schlichting,
942 and J. G. Kingsolver. 2014. Evolutionary change in continuous reaction norms. The American
943 Naturalist 183:453–467.

944 Nair, U. S. 1936. The standard error of Gini's mean difference. Biometrika 28:428–436.

945 Nakagawa, S., and R. Poulin. 2012. Meta-analytic insights into evolutionary ecology: an intro-
946 duction and synthesis. Evolutionary Ecology 26:1085–1099.

947 Nakagawa, S., R. Poulin, K. Mengersen, K. Reinhold, L. Engqvist, M. Lagisz, and A. M.
948 Senior. 2015. Meta-analysis of variation: ecological and evolutionary applications and beyond.
949 Methods in Ecology and Evolution 6:143–152.

950 Nakagawa, S., and E. S. A. Santos. 2012. Methodological issues and advances in biological
951 meta-analysis. Evolutionary Ecology 26:1253–1274.

952 O'Rourke, K. 2007. An historical perspective on meta-analysis: dealing quantitatively with
953 varying study results. Journal of the Royal Society of Medicine 100:579–582.

954 Plummer, M. 2010. JAGS version 2.0 Manual. International Agency for Research on Cancer.

955 Poissant, J., A. J. Wilson, and D. W. Coltman. 2010. Sex-specific genetic variance and the evo-
956 lution of sexual dimorphism: A systematic review of cross-sex genetic correlations. Evolution
957 64:97–107.

958 Postma, E. 2014. Four decades of estimating heritabilities in wild vertebrate populations:
959 Improved methods, more data, better estimates?, chap. 2, pages 16–33. Oxford University
960 Press.

961 Rosenberg, M. S. 2013. Moment and least-squares based approaches to meta-analytic inference,
962 chap. 9. Princeton University Press, Princeton, New Jersey.

963 Siepielski, A. M., J. D. DiBattista, and S. M. Carlson. 2009. It's about time: the temporal
964 dynamics of phenotypic selection in the wild. Ecology Letters 12:1261–1276.

965 Siepielski, A. M., K. M. Gotanda, M. B. Morrissey, S. E. Diamond, J. D. DiBattista, and S. M.
966 Carlson. 2013. The spatial patterns of directional phenotypic selection. Ecology Letters
967 16:1382–1392.

968 Sobel, J. M., and G. F. Chen. 2014. Unification of methods for estimating the strength of
969 reproductive isolation. Evolution 68:1511–1522.

970 Vetter, D., G. Rücker, and I. Storch. 2013. Meta-analysis: A need for well-defined usage in
971 ecology and conservation biology. Ecosphere 4:1–24.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 36:1–48.

Voje, K. L., T. F. Hansen, C. K. Egset, G. H. Bolstad, and C. Pelabon. 2014. Allometric constraints and the evolution of allometry. Evolution 68:866–885.

Walsh, B., and M. W. Blows. 2009. Abundant genetic variation + strong selection = multivariate genetic constraints: a geometric view of adaptation. Annual Review of Ecology, Evolution, and Systematics 40:41–59.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness-productivity relationship. Ecology 91:2522–2533.

Table 1: Formal meta-analysis of variation in selection gradients in the Kingsolver et al. (2001) dataset, stratified by broad (and informal) taxonomic groups, trait types, and fitness components.

dataset stratum	n	model 1 (using SEs)			model 2 (naive)		
		mean	SD	$E(\lvert\beta\rvert)$	mean	SD	$E(\lvert\beta\rvert)$
(a) broad taxonomic group							
vertebrates	218	0.016 (0.002 - 0.037)	0.086 (0.064 - 0.106)	0.071 (0.055 - 0.089)	0.035 (-0.008 - 0.068)	0.287 (0.263 - 0.315)	0.229 (0.211 - 0.253)
invertebrates	125	0.060 (0.012 - 0.099)	0.191 (0.146 - 0.231)	0.152 (0.124 - 0.193)	0.060 (0.014 - 0.112)	0.291 (0.255 - 0.327)	0.236 (0.208 - 0.268)
plants	62	0.031 (0.000 - 0.073)	0.133 (0.108 - 0.172)	0.117 (0.093 - 0.146)	0.008 (-0.056 - 0.058)	0.237 (0.193 - 0.275)	0.185 (0.156 - 0.221)
(b) trait type							
life history	49	0.007 (-0.021 - 0.025)	0.065 (0.036 - 0.089)	0.046 (0.031 - 0.072)	-0.013 (-0.044 - 0.023)	0.119 (0.098 - 0.147)	0.093 (0.078 - 0.118)
morphology	325	0.029 (0.007 - 0.051)	0.140 (0.113 - 0.157)	0.112 (0.094 - 0.129)	0.030 (-0.003 - 0.063)	0.300 (0.277 - 0.321)	0.242 (0.221 - 0.257)
(c) fitness component							
fecundity	55	0.074 (-0.003 - 0.135)	0.222 (0.164 - 0.282)	0.177 (0.138 - 0.232)	0.068 (-0.009 - 0.139)	0.281 (0.249 - 0.358)	0.242 (0.203 - 0.297)
mating success	109	0.091 (0.055 - 0.126)	0.131 (0.100 - 0.172)	0.129 (0.105 - 0.163)	0.100 (0.045 - 0.158)	0.298 (0.266 - 0.343)	0.263 (0.224 - 0.290)
survival	232	0.004 (-0.010 - 0.018)	0.060 (0.044 - 0.078)	0.048 (0.036 - 0.064)	-0.006 (-0.036 - 0.025)	0.249 (0.228 - 0.273)	0.198 (0.181 - 0.218)

Table 2: Mixed model-based estimates of variation in reaction norm intercepts, slopes, and curvatures. The main results are (a) variation in random coefficients among populations and (b) variation in random coefficients among populations, along with (c) residual variances of each of the four models with different standardisations of environmental variables and environment-specific mean phenotypes. Parts (d) and (e) report results from the same models, but transformed to represent mean absolute differences, rather than variances, and where the measures of curvature are re-scaled to second derivatives, rather than quadratic terms. Note that in parts (d) and (e) mean absolute differences are reported for second derivatives, which are twice the values of quadratic coefficients (and so their variance is four times that of the variance of quadratic coefficients), to allow comparison with metrics calculated in Murren et al. (2014).

	mean-standardised response		log response	
environmental standardisation:	equal range	equal interval	equal range	equal interval
(a) among-population variation (SD)				
intercept	0.179	0.174	0.121	0.083
slope	0.047	0.019	0.016	0.019
curvature	0.016	0.002	0.009	0.008
(b) among-species variation (SD)				
intercept	0.161	0.054	0.064	0.274
slope	0.071	0.093	0.011	0.146
curvature	0.061	0.010	0.001	0.039
(c) residual variation (SD)				
residual	0.288	0.306	0.307	0.309
(d) among-population mean absolute differences				
intercept	0.202	0.197	0.137	0.094
slope	0.053	0.021	0.018	0.022
second derivative	0.035	0.004	0.020	0.017
(e) among-species mean absolute differences				
intercept	0.182	0.061	0.285	0.309
slope	0.080	0.105	0.116	0.165
second derivative	0.134	0.023	0.071	0.090

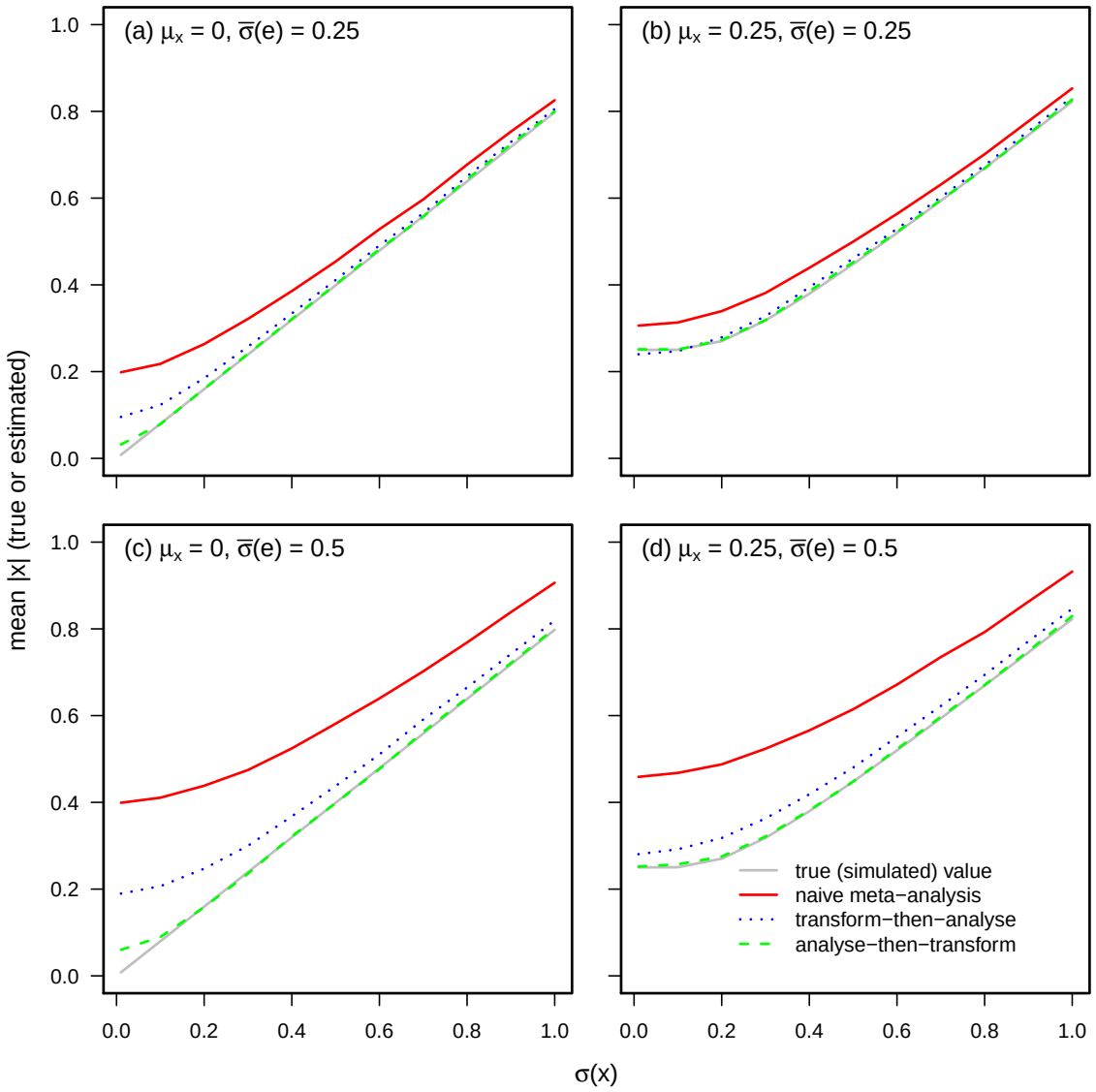

Figure 1: Bias in estimates of the mean absolute value of a meta-analytic quantity (x; all notation follows that given in the text) in three different approaches to meta-analysis. The different panels show results for different true mean values (μ_x) and mean standard errors ($\bar{\sigma}(e)$), and across a range of true standard deviations of the meta-analytic quantity ($\sigma(x)$). The 'transform-then-analyse' meta-analytic option calculates estimated absolute values and their standard errors, from the signed values and their standard errors in the meta-dataset, and then applies a random effects meta-analysis. The 'analyse-then-transform' option directly models the mean and variance of the (signed) values in the meta-dataset (accounting for their uncertainty via reported standard errors), and then obtains the mean absolute value from the inferred distribution of the original statistic.

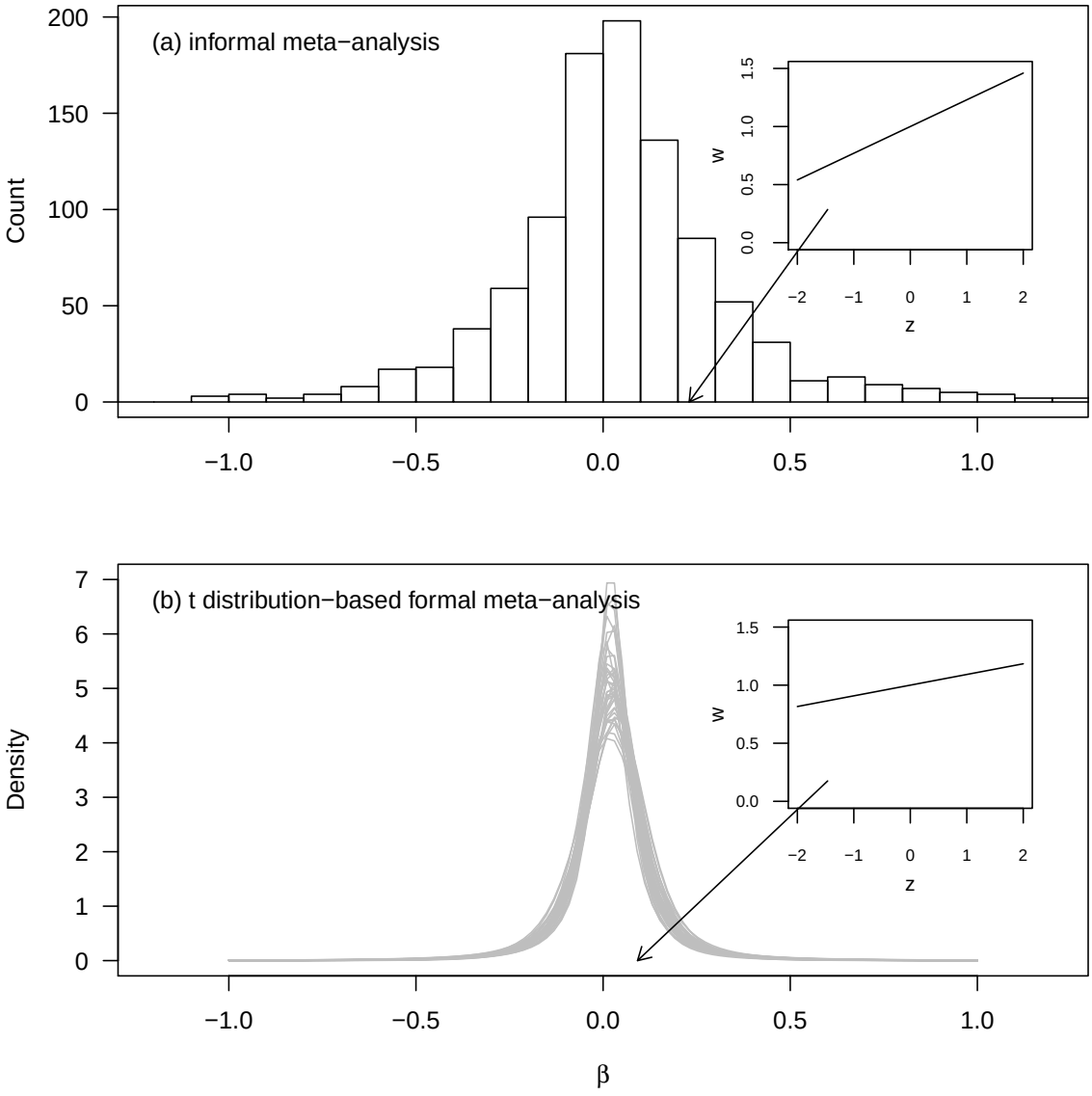

Figure 2: The distinction between distributions of estimated selection gradients and the distribution of selection. (a) the distribution of estimated directional selection gradients from the Kingsolver et al. (2001) meta-dataset. (b) 40 samples of the posterior distribution of a three parameter t-distribution based model estimating the distribution of directional selection gradients, accounting for the tendency for sampling error to inflate the apparent variation and mean magnitude (i.e., absolute value) of selection gradients. Inset plots depict the slopes of the relative fitness functions corresponding to the mean absolute value of selection gradients in each analysis.

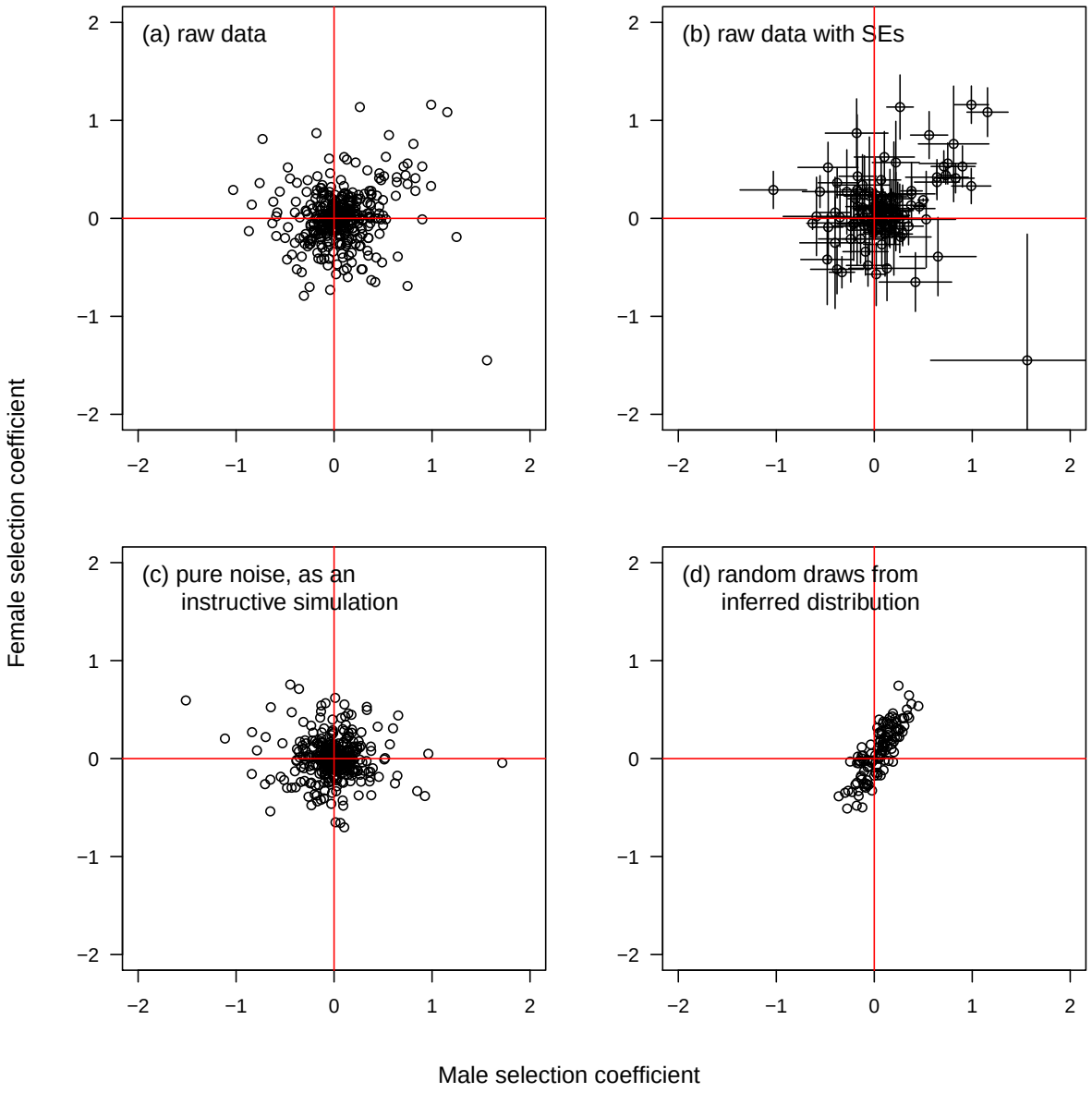

Figure 3: Observed and inferred distributions of male and female selection coefficients. (a) all data, (b) the subset of the data with available standard errors. (c) shows simulated pairs, where all male and female selection coefficients are zero, plus random noise drawn from the standard errors in the dataset. (d) shows random draws from a fitted model, accounting for sampling error.

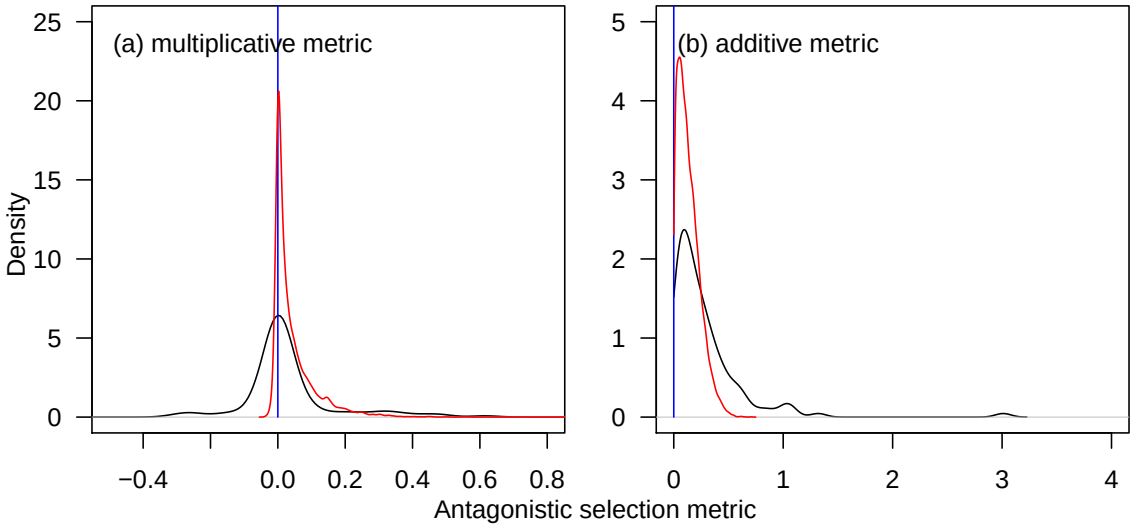

Figure 4: Distributions of two metrics of sexually-antagonistic selection, as applied to the raw data (black lines), and as inferred from a model that accounts for the effect of sampling error to bias inference of sexually-antagonistic selection (red lines). The multiplicative metric (a) is the product of male and female selection coefficients. Negative values occur when selection in males and females differs in sign, and positive values occur when the signs are the same across the sexes. Values near zero occur when there is little selection in one or both sexes. The additive metric (b) is the difference in male and female coefficients, and thus represents the distribution of total differences, but the values of the metric are not so directly tied to the coefficients in either sex.

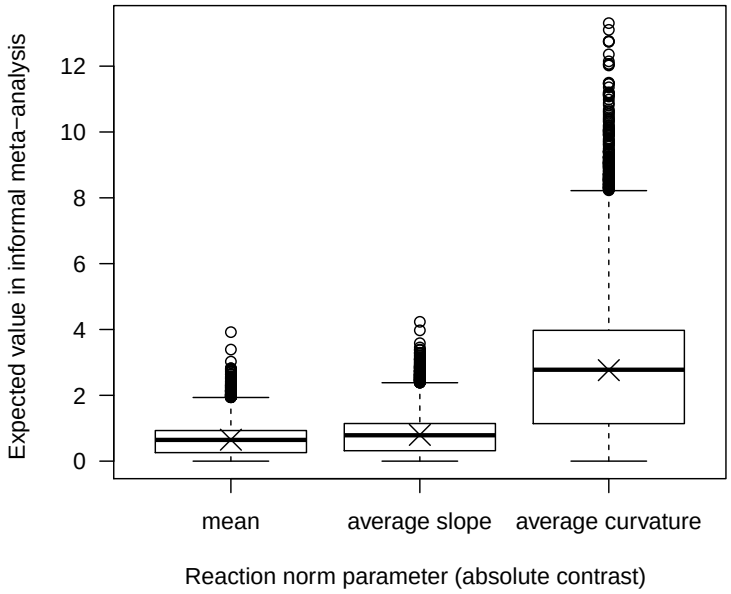

Figure 5: Analytical (crosses representing expectations) and simulation results (distributions in boxplots, with solid lines showing means) for bias in reaction norm parameters in an informal meta-analysis. For the special (and most frequent in the database) case of three environments, the analysis/simulation gives the expected values of the differences in average value, average slope, and average curvature between two reaction norms that are identical, but where residual variation exists in environment-specific estimated means. The case in this plot is for a residual variance of one unit, however this variance is arbitrary. The critical results are that (*i*) even in the limit of infinite data, the metrics do not converge on their true values (if zero, in this example), and (*ii*) the differences in the different metrics due to statistical noise alone follow a superficially interesting biological pattern.

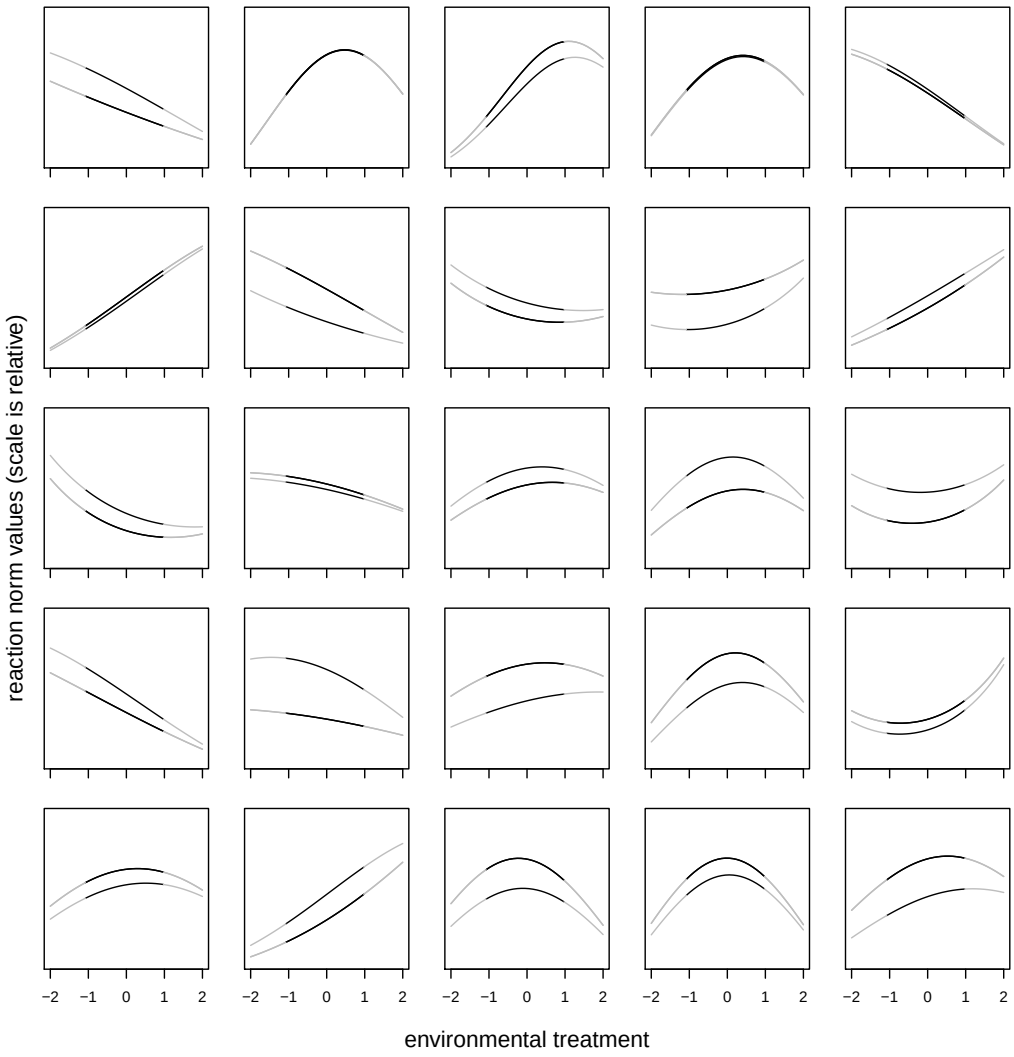

Figure 6: Simulated (log) quadratic approximations to intra-specific pairs of reaction norms, based on a random regression mixed-model analysis. The mixed-model analysis was conducted with the range of environmental variables in each study standardised to lie between -2 and +2. The values are somewhat arbitrary, and these specific values reflect loose assumptions that the relevant environmental variable might be normally-distributed in nature, and that researchers use their available resources to cover the majority of this range; under these assumptions, the scaling from -2 to +2 would make each unit equal to one SD of the environmental variable in nature. Quadratic approximations, or models of families of quadratic approximations, are most likely to provide good fits in the proportion of the range where the most data are available; the darker colouring from the environmental rage of -1 SD to +1 SD is arbitrary, but intended to draw focus to the range over which the model is likely to be most reliable.

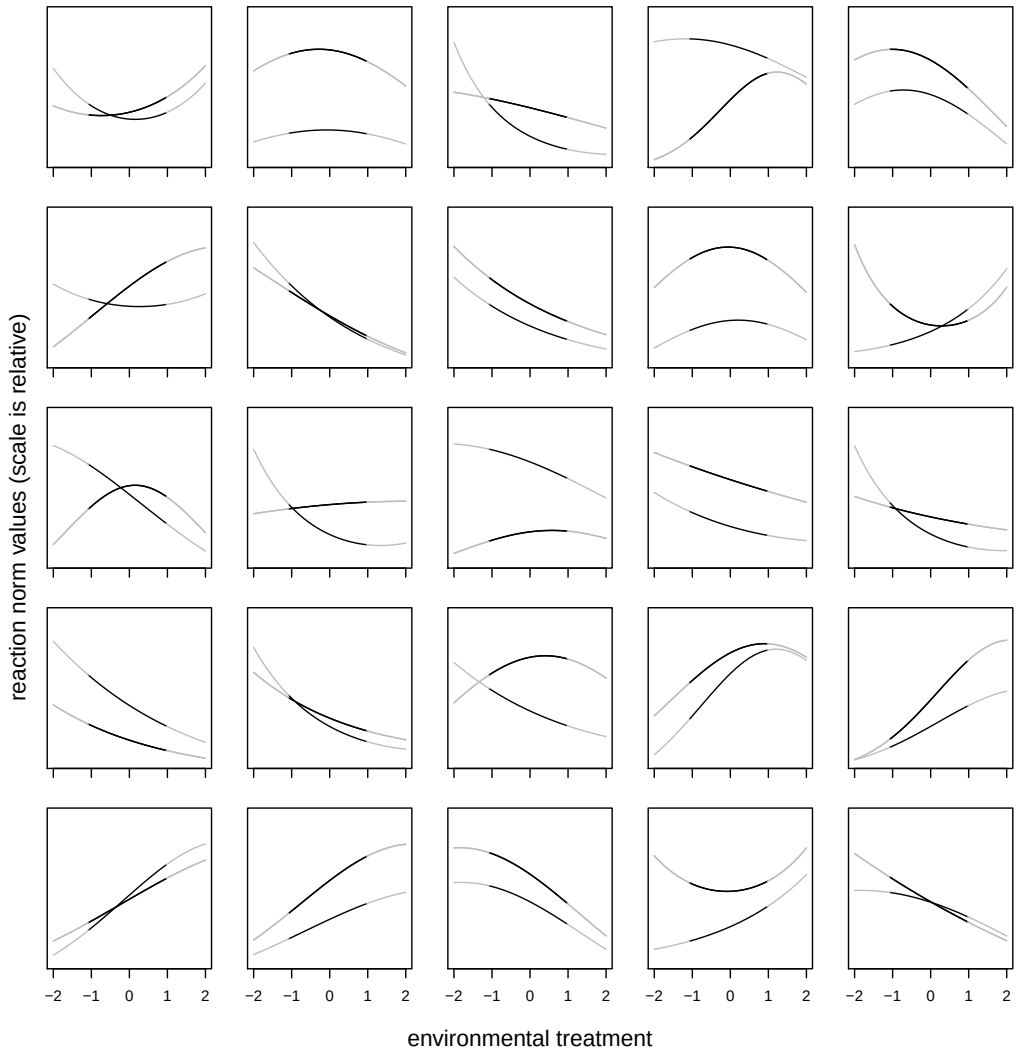

Figure 7: Simulated (log) quadratic approximations to congeneric pairs of reaction norms, based on a random regression mixed-model analysis. See text and caption of figure 6 for an explanation of the scaling and interpretation of the environmental variables.