Ecology Letters, (2003) **6**: 448–454

REPORT

Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques

Marc J. Lajeunesse and Mark R. Forbes*
Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, Canada, KIS 5B6
**Correspondence: E-mail: mforbes@ccs.carleton.ca*

Abstract

Variable reporting of results can influence quantitative reviews by limiting the number of studies for analysis, and thereby influencing both the type of analysis and the scope of the review. We performed a Monte Carlo simulation to determine statistical errors for three meta-analytical approaches and related how such errors were affected by numbers of constituent studies. Hedges' *d* and effect sizes based on item response theory (IRT) had similarly improved error rates with increasing numbers of studies when there was no true effect, but IRT was conservative when there was a true effect. Log response ratio had low precision for detecting null effects as a result of overestimation of effect sizes, but high ability to detect true effects, largely irrespective of number of studies. Traditional meta-analysis based on Hedges' *d* are preferred; however, quantitative reviews should use various methods in concert to improve representation and inferences from summaries of published data.

Keywords

Hedges' *d*, item response theory, Monte Carlo, publication bias, response ratio, type I error, type II error.

Ecology Letters (2003) 6: 448–454

INTRODUCTION

The comprehensive review of findings across studies is important to hypothesis testing and informing future research. Meta-analysis provides a quantitative method for reviewing studies and has become increasingly important for synthesizing research in ecology and evolution (e.g. Arnqvist & Wooster 1995a; Poulin 2000; Brown & Silk 2002;Moore & Wilson 2002;West & Sheldon 2002). One limitation of meta-analysis is its use and treatment of published studies with different analyses or approaches to reporting results. Some meta-analyses require that statistics used to measure the magnitude of relationships (known as effect sizes) are specified explicitly for each study. For instance, the commonly used metric for effect size, Hedges' *d*, restricts analysis to studies that have reported sample sizes, mean values and standard deviations or their surrogates (Hedges & Olkin 1985). However, it is not uncommon to find variable reporting of these basic statistics in ecological literature (Adams *et al.* 1997). This variable reporting is not only because of discrepancies in quality of research (Englund *et al.* 1999; Gurevitch & Hedges 2001), but also the result of editorial biases for brevity (Hunter & Schmidt 1990), or even reporting 'tastes'.

There is agreement that low quality studies should be excluded from analyses (i.e. studies that have flawed designs; Gurevitch & Hedges 1999). However, should all studies that do not fit precise requirements of statistical reporting be excluded from consideration? The aim of this paper is to address how variable reporting can affect numbers of studies included in analyses and the utility of various meta-analytical approaches.

Variable reporting of statistics can affect quantitative reviews in several ways. Overlooking studies that lack specific statistical information can decrease overall confidence in rejecting true null hypotheses (analogous to type I errors within studies, hereafter referred to as review type I errors or RTI). This problem is increased when moderator variables or covariates subdivide studies (Arnqvist & Wooster 1995b). In particular, researchers may require representation at few to several levels of moderator variables (e.g. taxa, latitude, trait type, and/or publication date). The impact of exclusion of studies failing to report basic statistics is largely unexplored, but can potentially affect hypothesis testing when the number of studies remaining is small. There is evidence that studies excluded in such a manner are not more variable or more aberrant than studies included (Englund *et al.* 1999).

It is also known that studies with low sample sizes contribute to larger error variances of effect sizes (also known as the funnel problem; see Light & Pillemer 1984). Many studies in ecology and evolution have low sample size, where low replication may be a consequence of study taxon, complexity of study, or animal care issues (Møller & Jennions 2001, 2002). Samples size limitations enlarge the probability of type II errors (accepting false null hypotheses). Several meta-analytical techniques attempt to control for *within study* type II errors. Researchers achieve this by correcting studies by their sample size, weighting studies during computation of mean values (to increase power of tests and precision of estimates; Gurevitch & Hedges 1999), and by not relying on actual statistical outcomes of studies (such as vote-count methods; Hedges & Olkin 1985). However, meta-analysis does not correct for similar errors at the review level, which is dependent on the number of studies under study (also known as second-order sampling error; Hunter & Schmidt 1990).

One way around second-order sampling error for quantitative reviews is to increase representation of studies. However, increased representation means relaxation of reporting criteria, and/or calls for access to unpublished research which increases representation and also addresses the file drawer problem associated with analysis on only the published subset of all studies (Arnqvist & Wooster 1995a;Møller & Jennions 2001). If there is low power within studies (regardless of whether there are reporting deficiencies), then increasing the number of studies used for the analysis may increase the power of observing an effect should it exist. Inclusion of more studies also allows stronger homogeneity tests across studies (testing if studies are of comparable effect sizes; Hedges & Olkin 1985). Homogeneity tests can help establish whether or not other moderator variables should be explored. Many important ecological and evolutionary questions may be addressed with alternatives to traditional meta-analyses (Englund *et al.* 1999). For instance, researchers interested in pattern description across several taxa run into problems of mixing measures quantified with different statistics. As we describe below, using alternatives to traditional meta-analyses can allow for these types of broad generalizations, and may be the only choice in instances where many studies lack complete statistical annotation.

A fine balance must be met between restricting analyses to studies where variance estimates are provided and obtaining an accurate estimate of between study variance by including several studies. There are techniques available to include studies if those studies fail to report measures of variance (see Gurevitch & Hedges 1999). However, such studies are treated less precisely (i.e. fewer statistics quantify each study). For instance, the log response ratio (RR) (also known as *lr* or RR) is a flexible measure that requires only mean values (Hedges *et al.* 1999), but can be expanded to include additional statistics depending on their availability (Gurevitch & Hedges 2001). Models based on item response theory (IRT) are also less restrictive, using only sample sizes and a categorical gauge of experimental outcome (Hedges & Olkin 1985). These gauges are not necessarily restricted to coarse dichotomous outcomes (e.g. significant or non-significant results, akin to vote-count methods; Lord 1980), but are flexible to fit particular research directions and/or various degrees of outcomes (e.g. significant effect, but not in the predicted direction; see van der Linden & Hambleton 1997). As there are fewer parameters needed, the estimate of effect size becomes coarser; thereby increasing type II error at the individual study level. However, analogous error rates at the review level (i.e. hereafter review type II or RTII errors) may decrease by including more studies.

Reviewers should base analyses on the largest number of studies available (Arnqvist & Wooster 1995b). If the collection of studies is largely incomplete, then conclusions drawn from analysis are limited in scope. The focus of this paper is to examine how variable reporting can affect the utility of meta-analytical approaches that differ in criteria for inclusion of constituent studies. We review the various methods to calculate effect size (Hedges' *d*, RR and IRT) and examine their susceptibility to RTI errors, or likelihood of falsely concluding effects. We also assess RTII errors, or likelihood of failing to detect true effects. We achieve this by using a Monte Carlo study simulating the two extremes of potential effects (i.e. effect and no-effect), and then compare error rates of different meta-analyses that are similar or different with respect to the number of constituent studies.

METHODS

Estimators of effect size

We used three methods to calculate effect sizes for experimental studies: Hedges' *d*, RR, and a model based on IRT. We briefly cover each method, but for more detailed summaries of these measures and their derivations see Hedges & Olkin (1985), Hedges *et al.* (1999), and van der Linden & Hambleton (1997).

Our first estimator of effect size is Hedges' *d*, which measures the difference between the control (c) and experimental (e) means in units of standard deviations, while correcting for small sample size bias (to avoid overestimating effect sizes when study sample size is low; Hedges & Olkin 1985). It combines for each *i*th experiment, the mean values ($\bar{Y}_i^c$ and $\bar{Y}_i^e$), standard deviations (s_i^c and s_i^e), and sample sizes (N_i^c and N_i^e) to give an effect size *d* calculated as follows:

$$d_i = \frac{\bar{Y}_i^{e} - \bar{Y}_i^{c}}{s_i}\left(1 - \frac{3}{4m_i - 1}\right), \quad (1)$$

where the pooled standard deviation $s_i = \{[(N_i^{e} - 1)(s_i^{e})^2 + (N_i^{c} - 1)(s_i^{c})^2]/m_i\}^{1/2}$ and $m_i = N_i^{e} + N_i^{c} - 2$ degrees of freedom. Variance of d_i is calculated following Gurevitch & Hedges (2001). Grand mean effect across studies with respective 95% confidence intervals (CI) are calculated using the sample size weighted method (also known as d_{++}; see Hedges & Olkin 1985). The advantage of using the weighted method is that studies with low sample sizes contribute less to effect size estimates.

The second metric used to estimate effect for each study is RR given by

$$RR_i = \ln\left(\bar{Y}_i^{e}/\bar{Y}_i^{c}\right) \quad (2)$$

Following Gurevitch & Hedges (2001), the grand mean effect across studies was calculated using the unweighted average method (with the weights of each study equalling one) and the nonparametric weighted method [weighted by $(N_i^{e}N_i^{c})/(N_i^{e} + N_i^{c})$; see Gurevitch & Hedges 2001]. There is also the parametric method for weighting studies (which weight studies by sample sizes and standard deviations; see Hedges *et al.* 1999). However, we do not include a comparison of parametric RR because it has similar inclusion criteria as Hedges' *d*. The aim of our paper is to compare methods differentially influenced by variation in statistical reporting, thereby affecting the number of studies for analysis and thus affecting review level error rates. We focus our comparison with the unweighted RR (hereafter RR_u) and the nonparametric weighted RR (or RR_w), as these approaches are the least restrictive alternatives to Hedges' *d*. Analytical methods are not available to calculate variance for grand mean RR_u and RR_w. Thus the 95% CI around grand means were bootstrapped using the bias correction for small samples method (with 4999 iterations; Adams *et al.* 1997; Dixon 2001).

Our final measure is based on the normal ogive model of IRT (Hedges & Olkin 1985), which requires samples sizes and a response outcome (X_i) for each study. Responses can be any number of different research outcomes (van der Linden & Hambleton 1997), but for our purposes we assume that X_i equals one, if a statistic testing differences between $\bar{Y}_i^{c}$ and $\bar{Y}_i^{e}$ (e.g., Student's *t* test) exceeds the critical value for that statistic (at the 0.05 significance level), whereas X_i equals zero for all other cases. The grand effect across k studies is evaluated with

$$L(\delta|X_1,\ldots,X_k) = \sum_{k} X_i \log[1 - \Phi(-\sqrt{\tilde{n}_i}\delta)] + \sum_{k} (1 - X_i)\log[\Phi(-\sqrt{\tilde{n}_i}\delta)], \quad (3)$$

where studies are weighted by $\tilde{n}_i = (N_i^{e}N_i^{c})/(N_i^{e} + N_i^{c})$ and Φ is the standard normal distribution function. Iterating eqn 3 for an array of δ effect values, we picked the largest δ as the effect size ($\hat{\boldsymbol{\delta}}$). A limitation of IRT is that it is unable to directly estimate $\hat{\boldsymbol{\delta}}$ if all studies have similar outcomes (i.e. effect sizes tend towards $+\infty$ if all X_i are one, or towards $-\infty$ if all zero; Lord 1980). However, a maximum effect size ($\hat{\boldsymbol{\delta}}_{MAX}$) can be estimated indirectly by summing the weights of effects for each study. These weights were obtained by sequentially modifying the research outcome of each study and re-estimating its grand effect size for the jth group ($\hat{\boldsymbol{\delta}}_j$; e.g., if $X_k = \{1, 1, 1\}$ then $\hat{\boldsymbol{\delta}}_1$ was estimated for $\{0, 1, 1\}$, $\hat{\boldsymbol{\delta}}_2$ for $\{1, 0, 1\}$ and $\hat{\boldsymbol{\delta}}_3$ for $\{1, 1, 0\}$). Tabulating these results in the estimated effect column $\tilde{\boldsymbol{\delta}}_k = \{\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_j\}$ and solving the system of linear equations $\mathcal{A}_{x,y}\bar{\boldsymbol{\delta}}_j = \tilde{\boldsymbol{\delta}}_k$ for $\bar{\boldsymbol{\delta}}_j$, we get $\hat{\boldsymbol{\delta}}_{MAX}$ by summing the solved study weights in $\bar{\boldsymbol{\delta}}_j$. $\mathcal{A}_{x,y}$ is a symmetric matrix (where x and y equal k number of studies) containing -1 when $x = y$ (indicating the change of study outcome) and 1 when $x \neq y$ (indicating no change). Variance and 95% CI were estimated using the large sample variance method (Hedges & Olkin 1985).

Monte Carlo study

Our first approach is to construct plausible 'studies' with sample sizes, mean values, and standard deviations as seen in real reviews, followed by a Monte Carlo simulation to estimate the RTI and RTII error rates of each approach to measuring effect sizes. In our first instance, the true effect in the simulated data was set to zero (i.e. H_0true, $\bar{Y}_i^{c} = \bar{Y}_i^{e}$), whereas in the second instance the effect was set larger than zero (i.e. H_0false, $\bar{Y}_i^{c} \neq \bar{Y}_i^{e}$). Our simulation used the study as the unit of analysis to avoid several review pitfalls; such as non-independence, inflated sample sizes as a result of extracting several effect sizes from single studies, and/or the 'apples-and-oranges' problem (averaging effects across dependent and independent variables; see Hunter & Schmidt 1990). We also consider instances where some of the studies had incomplete annotation of results such that they could not be included in calculations of Hedges' *d* (e.g. standard deviations were not reported).

We begin by randomly generating sample sizes for controlled and experimental groups of each study, where $N_i^{c} \sim \mathbf{N}(15, 1)$ for $i = 1,2,\ldots,k$ and $N_i^{e} \sim [N_i^{c} + \mathbf{N}(0, 1/2)]$ for $i = 1,2,\ldots,k$. The $\sim$ denotes that samples sizes for a group of k studies are distributed normally ($\mathbf{N}$) around a mean of 15 with an error structure of one standard deviation (as the case for N_i^{c}). We generated studies with low sample sizes because these are common to ecological and evolutionary studies (see Møller & Jennions 2002), and correlated N_i^{e} to N_i^{c} as these are often similar because of logistical reasons of experimentation within studies.

With these sample sizes, we first simulate a group of studies where the true effect was set to zero (no effect). 'No effect' data for each study (i.e. mean values and standard deviations) were calculated from replications generated as follows: $Y^{c}_{i,j} \sim \mathbf{N}[50, (N^{c}_{i})^{-1/2}]$ for $j = 1, 2, \ldots, N^{c}_{i}$ and $Y^{e}_{i,j} \sim \mathbf{N}[50, (N^{e}_{i})^{-1/2}]$ for $j = 1, 2, \ldots, N^{e}_{i}$. These data were generated such that sampling error was the only artefact that produced false variation across studies; where the sizes of departures from the true mean (in this case 50) is dependent on sample sizes for both $\bar{Y}^{c}_{i}$ and $\bar{Y}^{e}_{i}$. Any difference between $\bar{Y}^{c}_{i}$ and $\bar{Y}^{e}_{i}$ is thus the result of sampling error. Our second data set simulated studies with true effect size differences between control and experimental groups. Again, we generated data from random samples (as above), but in this scenario experimental group mean values were distributed outside the upper 95% CI of the control mean values (following Zar 1984), specifically $Y^{c}_{i,j} \sim \mathbf{N}[50, (N^{c}_{i})^{-1/2}]$ for $j = 1, 2, \ldots, N^{c}_{i}$ and $Y^{e}_{i,j} \sim \mathbf{N}\{50 + 2[t_{(0.05,N-1)}N^{-1}], (N^{e}_{i})^{-1/2}\}$ for $j = 1, 2, \ldots, N^{e}_{i}$, and where N is the true population sample size (15 replications; see above) and $t_{(0.05,N-1)}$ is the inverse of the Student's t-distribution for N–1 degrees of freedom. Here, any similarities between $\bar{Y}^{c}_{i}$ and $\bar{Y}^{e}_{i}$ are only the result of sampling error, and the direction of effect should be positive (experimental differences are larger than controls). With these data, effect sizes were calculated with Hedges' d and RR. IRT requires an additional statistic to quantify differences between each mean (e.g. experimental outcome X_i; see above). We used t-tests with unequal sample size to gauge these mean differences (following Zar 1984).

Each study had samples sizes, mean values and standard deviations of control and experimental groups and an outcome of a t-test statistic. One thousand studies for each 'no-effect' and 'effect' groups were simulated. From these studies, 5, 10, 15, 20, 25 and 30 test studies were randomly sampled (1000 times with replacement) and had their grand mean effect size estimated using the three meta-analytical methods outlined above (with RR including both unweighted and weighted approaches). Such small sample sizes are warranted particularly for meta-analysis with moderator variables. For all cases, we measured whether the 95% CI of the grand mean effect size overlapped with zero (detecting no effect), or did not overlap with zero (detecting an effect). The probability for committing an RTI error is estimated as the fraction of tests that erroneously rejected the null hypothesis (from the 'no-effect' group of studies), and RTII errors were estimated as the fraction that erroneously accepted the null hypothesis (from the 'effect' group of studies). Because of the way the studies were generated in the 'effect' group, significant studies are more likely to be sampled than non-significant studies. This sampling bias is a problem for Hedges' d, as it is known to overestimate effect sizes when only significant studies are observed (Hedges & Olkin 1985). We corrected the bias in Hedges' d using the methods outlined in Hedges (1984).

We also note that the appropriate comparison between effect size metrics is not with respect to the same numbers of constituent studies, but rather a comparison of RR and IRT based on more studies and Hedges' d based on fewer, perhaps far fewer, studies. This comparison reflects 'reviews' in which reporting criteria to calculate Hedges' d were underrepresented to varying degrees (i.e. because of missing standard deviations).

RESULTS

Review type I error rates

The RTI error rates for effect sizes calculated by all three methods, indicated by the number of times 'an effect' was detected when in fact it did not exist, decreased as the number of studies increased (Table 1). Permissible RTI error rates are found below the 0.05 α-value (Zar 1984). Hedges' d and IRT had similar error rates, each showing a substantial increase in accuracy with at least 15 studies (Table 1), and achieved acceptable error rates at 15 and 16 studies, respectively (RTI for IRT at 16 studies was 0.031). RR_u and RR_w had lower error rates than either Hedges' d and IRT when 10 or fewer studies were analysed (Table 1); however, RR_u and RR_w still maintained error rates above 0.13 with 30 studies (Table 1; weighting RR by sample sizes of constituent studies only slightly improved error rates). This relatively high likelihood of committing an RTI error for approaches based on RR implies that this technique may overestimate effects, which will have implications for assessing RTII errors. Our next comparisons are based on the supposition that Hedges' d would have analysed fewer studies than either RR or IRT (because of excluding studies with incomplete statistical annotation). We found that when review sample sizes are low for Hedges' d (e.g. <15 studies), a complementary analysis with IRT (which would include studies excluded by Hedges' d) could increase the confidence of conclusions drawn from analyses. IRT can achieve acceptable RTI error rates (below 0.05) when analyses are based on 12.5–68.8% additional studies, otherwise excluded by Hedges' d. These results are not shown, but can be easily extrapolated from Table 1 where Hedges' d is based on five to fourteen of 16 available studies such that $2/16 = 12.5\%$ additional studies.

Review type II error rates

Again, precision of Hedges' d and IRT metrics increased in relation to the number of studies analysed (Table 1). Hedges' d had much lower error rates than IRT for detecting true effect differences. IRT error rates gradually

Table 1 Monte Carlo simulation of the probability of review type I errors (RTI; finding differences when there are none) and review type II errors (RTII; failure to detect true effects) of grand mean effect sizes

	RTI error				RTII error			
k	d	IRT	RR_u	RR_w	d	IRT	RR_u	RR_w
5*	0.710	1.000	0.281	0.266	0.885	1.000	0.006	0.003
10	0.233	0.339	0.191	0.181	0.410	0.769	0.001	0.000
15	0.033	0.096	0.178	0.170	0.043	0.582	0.000	0.000
20	0.003	0.004	0.164	0.152	0.001	0.475	0.000	0.000
25	0.000	0.001	0.162	0.150	0.000	0.349	0.000	0.000
30	0.000	0.000	0.159	0.134	0.000	0.245	0.000	0.000

*IRT is unable to effectively estimate a gross effect size for $k \leq 5$ (Lord 1980).
Grand mean effect sizes are estimated by Hedges' d (d), item response theory (IRT), and unweighted and weighted log response ratio (RR_u and RR_w, respectively). The number of constituent studies sampled (k) from 1000 studies designed show no treatment effects (no differences between experimental and control groups), or show treatment effects (with true differences between experimental and control groups). RTI errors are based on the number of times 95% confidence intervals (CI) of the grand mean effect size did not overlap with zero effects, and RTII error on the number of times 95% CI did overlap with zero effects.

Table 2 The IRT susceptibility to within study type II errors (finding no effect when there is one; see Table 1)

	X_k	N	M	IQR	R
N^{h} *	n.s.	252	10.45	7.91–13.08	4.36–19.95
	<0.05	748	16.36	13.04–19.94	5.45–32.59
Power of t-test†	n.s.	252	0.406	0.199–0.634	0.013–0.970
	<0.05	748	0.849	0.514–0.994	0.003–1.000

*Kruskal–Wallis $\chi^2 = 246.4$, d.f. $= 1$, $P < 0.001$.
†Kruskal–Wallis $\chi^2 = 179.4$, d.f. $= 1$, $P < 0.001$.
Comparison of harmonic mean values of experimental and control sample sizes, $N^{h} = (2N^{e}N^{c})/(N^{e} + N^{c})$, and power of unequal two-tailed t-tests (as calculated in Zar 1984, p. 136) for groups categorized (X_k) as either above or below critical value of t-test at the 0.05 level. Medians (M) and inter-quartile and total ranges (IQR and R) are derived from 1000 randomly generated experiments (N) designed to have differences between experimental and control groups (see Methods)

decreased as the number of studies increased, but still retained an error above *c.* 0.24 (Table 1). IRT has difficulty in detecting true effects when within study type II errors are high (despite weighting studies by their sample size). Low precision of IRT to detect effects was the result of the coarse categorization of research outcomes based on statistical testing (Table 2), where many studies were categorized as non-significant when having been assigned low sample sizes randomly, thereby resulting in low power (Table 2). These studies biased RTII error rates towards falsely rejecting true effects (Table 1).

Note also that our RTII data set was constructed such that studies with null sample effects occurred. Inclusion of these studies should affect error rates when review sample sizes are low. RTII error rates of Hedges' d were sensitive to these studies (Table 1), however both RR_u and RR_w showed the least susceptibility to RTII errors (Table 1), where any analyses above five studies nearly always detected effects. However, these error rates may be misleading. Unweighted results overestimate actual effects (inflating true null effects when they exist; Gurevitch and Hedges 1999), and bootstrapped 95% CI are biased against detecting null effects when the direction of all study outcomes are in the same, non-zero direction as can occur randomly (Adams *et al.* 1997). The high RTI error rates of RR may give an indication of its propensity to reject null effects (Table 1), which are further reflected in low RTII error rates.

DISCUSSION

We designed our Monte Carlo simulation to explore the susceptibility of three meta-analytical procedures to errors at the review level analogous to type I and type II errors at the study level. The factors causing variation between studies were sampling errors and variation in reporting of statistics. We found that as the number of studies increased, RTI and

RTII errors decreased for the majority of meta-analytic procedures. Specifically, Hedges' *d* and IRT performed similarly against RTI errors, but IRT had difficulties in detecting effects when present (high RTII error rates). Error rates of RR_u and RR_w was slightly affected by the number of studies analysed; however, they generally had either low precision for detecting null effects or high ability to detect true treatment effects.

With specific reference to RTI errors, we found that the efficacy of Hedges' *d* was affected by exclusion of studies. In a recent meta-analysis, many more studies met the criteria for IRT than for Hedges' *d* leading to use of the former method (Lajeunesse & Forbes 2002). In Van Zandt & Mopper's (1998) study, 70% more studies met the criteria for RR than for Hedges' *d*; however, both methods were used (see below). Earlier work combined with this study suggests that IRT could have improved error rates than Hedges' *d*, with respect to how RTI errors decrease with additional studies (see Table 1).

When accounting for RTII errors, Hedges' *d* retained greater ability to detect effects than IRT, even when IRT had analysed 30% more studies than Hedges' *d* (Table 1). Both RR_u and RR_w appeared to have the greatest ability to detect true effects. However, such low error rates may be the result of the general propensity for RR to conclude effects when they do not occur (high RTI error rates; Table 1). Suggested acceptable RTII error rates for meta-analysis range from 0.2 to 0.25 (Cohen 1977; Hunter & Schmidt 1990). IRT did meet the higher of these two acceptable rates, but only at 30 studies. IRT is inferior to Hedges' *d* or RR in detecting differences between test populations, because of constituent studies having low power and IRT being based partly on statistical outcomes (see Table 2). It is relevant that Hedges' *d* needs to be corrected as it overestimates effect sizes when significant studies are more common than non-significant studies (Hedges' *d* is monotonically related to *t*-tests; Hedges & Olkin 1985). Similar statistical corrections have yet to be developed for RR and IRT.

The choice of an effect size metric is a decision of critical importance (Osenberg *et al.* 1999). Each meta-analytical procedure has merits and faults. RTI and RTII error rates are two criteria that should affect choice of an effect size metric. In summary, Hedges' *d* has high within study precision (i.e. use of multiple statistics to quantify each study), yet these requirements also restrict inclusion of studies (which can be a problem with small review sample sizes). RR can be scaled to fit additional statistics, and is the most flexible of all procedures, but may require randomization procedures to estimate error variance, which makes the two sources of variance difficult to separate (Gurevitch & Hedges 1999). Analytical methods are available for calculating variance for RR, but require a measure of variance within each study (i.e. standard deviations), thereby further narrowing its inclusion criteria (see Hedges *et al.* 1999). IRT has simple requirements, but requires technical understanding, and is greatly affected by the type of measurement used to categorize each study (particularly true if these are coarsely gauged as significant/non-significant outcomes). These difficulties can be overcome by modifying IRT models to include more continuous categories of experimental outcomes (e.g. partial-credit models; van der Linden & Hambleton 1997). The utility of IRT is further augmented by its ability to include studies with different statistical approaches (e.g. reporting of nonparametric vs. parametric statistics). Weighted and unweighted RR have high RTI error rates and IRT has high RTII error rates. These effect size measures should perhaps be treated as analytical tools for directing research rather than viewed as hypothesis testing tools *per se*.

Use of these meta-analytical procedures is dependent on the goal of the review: to make conclusions specific to the studies reviewed, or to determine population parameters and 'true' experimental effects. However, reviewers need not be restricted to a single approach. For instance, Van Zandt & Mopper (1998) corroborated results when testing for local adaptation in herbivorous insects. They used 10 studies where effect sizes could be calculated as Hedges' *d*, but also used weighted RR based on 17 studies. Using RR enabled a broader review of potential moderator variables affecting patterns of local adaptation in these insects. Also, this review examined whether conclusions drawn from across studies were sensitive to the inclusion of a study with a high effect size. Such sensitivity analyses are important when results are based on small review sample sizes. Recent development of meta-analytical packages like *MetaWin* allow reviewers to analyse studies with different effect size metrics (in addition to providing bootstrapping procedures; Rosenberg *et al.* 1997, 2000). Using more studies would allow greater confidence for reviewers to draw inferences from analyses. The degree of concordance between two or more analyses could be informative regarding the strength of effect, or whether an artefact may have been included in less inclusive analysis (cf. Hunter & Schmidt 1990).

The results of this study show how variable reporting can affect the confidence of review findings. We find that conclusions in reviews may be affected by the type approach used to synthesise studies, but are more affected by the number of studies analysed, and problems within studies (incomplete statistical annotation, low sample size, etc.). It is known that studies with null results are less likely to be published and are less likely to report all statistics (Hunter & Schmidt 1990;Arnqvist & Wooster 1995a;Møller & Jennions 2001). Exclusion of inadequately reported studies may further bias analyses of published studies by contributing to the file drawer problem (see Møller & Jennions 2001).

Methods have been developed to quantify the number of unpublished/unlocated studies of null effect (see 'fail safe *N*' calculations; Rosenthal 1991). However, tests for assessing the effects of variable reporting have yet to be developed. A common suggestion to remedy this problem is for authors and editors of journals to be more rigorous in reporting study results (Gurevitch & Hedges 1999). However, many reviews are also based on historical data (using studies based in different reporting practices and/or different editorial requirements), particularly those determining whether effects show a temporal component. These studies may have important merits such as providing, by their inclusion, greater representation across moderator variables of interest. Researchers preferring more rigorous approaches should routinely give an indication of how many studies were excluded, as a result of not fitting requirements of specific meta-analyses. It is important to understand the numbers of studies (un)available for analysis, or the degree to which certain variables of interest are underrepresented to inform future research. Such information will help researchers address 'gaps' in the literature (cf. Lajeunesse & Forbes 2002) or perhaps influence policy on reporting standards.

ACKNOWLEDGEMENTS

We greatly appreciated the helpful comments of three anonymous reviewers. This study was funded by an NSERC grant to M.R.F.

REFERENCES

Adams, D.C., Gurevitch, J. & Rosenberg, M.S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology*, 78, 1277–1283.

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

Arnqvist, G. & Wooster, D. (1995b). Reply from G. Arnqvist and D. Wooster. *Trends. Ecol. Evol.*, 10, 460–461.

Brown, G.R. & Silk, J.B. (2002). Reconsidering the null hypothesis: is maternal rank associated with birth sex ratios in primate groups? *Proc. Nat. Acad. Sci. USA*, 99, 11252–11255.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences (Revised Edition)*. Academic Press, New York.

Dixon, A. (2001). Bootstrapping and randomization. In: *Design and Analysis of Ecological Experiments*, 2nd edn (ed. Scheiner, S.M. & Gurevitch, J.). Oxford University Press, Oxford, pp. 267–289.

Englund, G., Sarnell, O. & Cooper, S.D. (1999). The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology*, 80, 1132–1141.

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

Gurevitch, J. & Hedges, L.V. (2001). Meta-analysis: combining results of independent experiments. In: *Design and Analysis of Ecological Experiments*, 2nd edn (ed. Scheiner, S.M. & Gurevitch, J.). Oxford University Press, Oxford, pp. 347–369.

Hedges, L.V. (1984). Estimation of effect sizes under non-random sampling: the effects of censoring studies yielding statistically insignificant mean differences. *J. Edu. Stat.*, 9, 61–85.

Hedges, L.V., Gurevitch, J. & Curtis, P.S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, 80, 1150–1156.

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage Publications, Beverly Hills, CA.

Lajeunesse, M.J. & Forbes, M.R. (2002). Host range and local parasite adaptation. *Proc. R. Soc. Lond. B*, 269, 703–710.

Light, R.J. & Pillemer, D.B. (1984). *Summing Up: The Science of Reviewing Research*. Harvard University Press, Cambridge.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Moore, S.L. & Wilson, K. (2002). Parasites as a viability cost of sexual selection in natural populations of mammals. *Science*, 297, 2015–2018.

Møller, A.P. & Jennions, M.D. (2001). Testing and adjusting for publication bias. *Trends Ecol. Evol.*, 16, 580–586.

Møller, A.P. & Jennions, M.D. (2002). How much variance can be explained by ecologists and evolutionary biologists? *Oecologia*, 132, 492–500.

Osenberg, C.W., Sarnelle, O., Cooper, S.D. & Holt, R.D. (1999). Resolving ecological questions through meta-analysis: goals, metrics and models. *Ecology*, 80, 1105–1117.

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? *Proc. R. Soc. Lond.* B, 267, 787–792.

Rosenberg, M.S., Adams, D.C. & Gurevitch, J. (1997). *MetaWin: statistical software for meta-analysis with resampling tests*. Sinauer Associates, Sunderland.

Rosenberg, M.S., Adams, D.C. & Gurevitvh, J. (2000). *MetaWin 2.0: statistical software for meta-analysis*. Sinauer Associates, Sunderland.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Sage, Newbury Park.

Van Zandt, P.A. & Mopper, S. (1998). A meta-analysis of adaptive deme formation in phytophagous insect populations. *Am. Nat.*, 152, 595–604.

West, S.A. & Sheldon, B.C. (2002). Constraints in the evolution of sex ratio adjustment. *Science*, 295, 1685–1688.

Zar, J.H. (1984). *Biostatistical Analysis*. Prentice Hall, Englewood Cliffs.

van der Linden, W.J. & Hambleton, R.K. (1997). *Handbook of modern item response theory*. Springer-Verlag, New York.

Manuscript received 4 December 2002
First decision made 7 January 2003
Second decision made 13 February 2003
Accepted for publication 17 February 2003