Biol. Rev. (2007), **82**, pp. 591–605.
doi:10.1111/j.1469-185X.2007.00027.x

Effect size, confidence interval and statistical significance: a practical guide for biologists

Shinichi Nakagawa[1,*] and Innes C. Cuthill[2]

[1] *Department of Animal and Plant Sciences, University of Sheffield, Sheffield S*10 2*TN, UK* (E-mail: itchyshin@yahoo.co.nz)
[2] *School of Biological Sciences, University of Bristol, Bristol BS*8 1*UG, UK* (E-mail: i.cuthill@bristol.ac.uk)

(*Received* 2 *January* 2007; *revised* 24 *July* 2007; *accepted* 27 *July* 2007)

ABSTRACT

Null hypothesis significance testing (NHST) is the dominant statistical approach in biology, although it has many, frequently unappreciated, problems. Most importantly, NHST does not provide us with two crucial pieces of information: (1) the magnitude of an effect of interest, and (2) the precision of the estimate of the magnitude of that effect. All biologists should be ultimately interested in biological importance, which may be assessed using the magnitude of an effect, but not its statistical significance. Therefore, we advocate presentation of measures of the magnitude of effects (i.e. effect size statistics) and their confidence intervals (CIs) in all biological journals. Combined use of an effect size and its CIs enables one to assess the relationships within data more effectively than the use of *p* values, regardless of statistical significance. In addition, routine presentation of effect sizes will encourage researchers to view their results in the context of previous research and facilitate the incorporation of results into future meta-analysis, which has been increasingly used as the standard method of quantitative review in biology. In this article, we extensively discuss two dimensionless (and thus standardised) classes of effect size statistics: *d* statistics (standardised mean difference) and *r* statistics (correlation coefficient), because these can be calculated from almost all study designs and also because their calculations are essential for meta-analysis. However, our focus on these standardised effect size statistics does not mean unstandardised effect size statistics (e.g. mean difference and regression coefficient) are less important. We provide potential solutions for four main technical problems researchers may encounter when calculating effect size and CIs: (1) when covariates exist, (2) when bias in estimating effect size is possible, (3) when data have non-normal error structure and/or variances, and (4) when data are non-independent. Although interpretations of effect sizes are often difficult, we provide some pointers to help researchers. This paper serves both as a beginner's instruction manual and a stimulus for changing statistical practice for the better in the biological sciences.

Key words: Bonferroni correction, confidence interval, effect size, effect statistic, meta-analysis, null hypothesis significance testing, *p* value, power analysis, statistical significance.

CONTENTS

* Address for correspondence: Tel: +44114 222 0113; Fax: +44114 222 0002. E-mail: itchyshin@yahoo.co.nz

I. INTRODUCTION

The statistical approach commonly used in most biological disciplines is based on null hypothesis significance testing (NHST). However, the NHST-centric approach is rare amongst mathematically trained statisticians today and is becoming marginalised in biomedical statistics (particularly in the analysis of clinical drug trials), psychology and several other social sciences (Wilkinson & the Task Force on Statistical Inference, 1999; Altman *et al.*, 2001; American Psychological Association, 2001; Kline, 2004; Fidler *et al.*, 2004; Grissom & Kim, 2005). It is also the centre of current debate and imminent change in some areas of ecology and conservation science (Stephens *et al.*, 2005; Fidler *et al.*, 2006; McCarthy, 2007; Stephens, Buskirk & Del Rio, 2007). These movements are not surprising since NHST does not provide us with what are probably the two most important pieces of information in statistical inference: estimates of (1) the magnitude of an effect of interest (or a parameter of interest) and (2) the precision of that estimate (e.g. confidence intervals for effect size). NHST only informs us of the probability of the observed or more extreme data given that the null hypothesis is true, i.e. p value, upon which we make a dichotomous decision: reject or fail to reject. This paper explains how NHST misleads, why the presentation of unstandardised and/or standardised effect sizes and their associated confidence intervals (CIs) is preferable, and gives guidance on how to calculate them. We feel that it is the absence of accessible recommendations and systematic guidelines for effect size presentation, as much as an ignorance of the issues, which has hindered the spread of good statistical practice in the biological literature (e.g. Nakagawa, 2004; Nakagawa & Foster, 2004; Garamszegi, 2006).

II. WHY DO WE NEED EFFECT SIZE?

(1) Null hypothesis significance testing misleads

We will not provide a comprehensive list of the problems of NHST and associated p value here; this has already appeared elsewhere (Harlow, Mulaik & Steiger, 1997; Nickerson, 2000; Kline, 2004). Instead, we describe the three problems which we consider most relevant to the biological sciences.

First, in the real world, the null hypothesis can rarely be true. We do not mean that NHST can only reject, or fail to reject, rather than support the null hypothesis; rather that the null hypothesis itself is usually false. Consider a nominally monomorphic species of bird. Measuring the wing lengths of a large sample of males and females (say 1000 individuals) yields no significant sex difference and the researcher, well trained in classical statistics, concludes that the null hypothesis cannot be rejected. However, if one could somehow measure every single male and female in the species (i.e. the population that the sample of 1000 individuals was used to draw inferences about), then there would unquestionably be a difference in the mean wing length of males and females. If no sex difference was evident, this would only be due to a lack of measurement precision (e.g. the means may be identical to the nearest 0.1 mm, but not to the nearest 0.00001 mm). The only instance in which the null hypothesis may be exactly true is for categorical data; for example the sex ratio (number of males and females in a population) may indeed be exactly equal, but this is likely to be a transient and infrequent state of affairs. Of course what matters in the case of wing length or sex ratio is that the difference is too small to be biologically important, but this is a matter of biological inference, not statistics; the null hypothesis itself cannot be true (nor is it biologically relevant whether it is exactly true).

Second, NHST and the associated p value give undue importance to just one of the hypotheses with which the data may be consistent. To understand why this may be misleading, it is useful to consider what is sometimes termed the counter-null hypothesis (Rosenthal, Rosnow & Rubin, 2000). As a simple example, consider a measured change in some continuous variable (Fig. 1). The mean change is 10 units but, with the observed variation, the 95% confidence intervals include zero (say -1 to $+21$). A one-sample t-test is therefore non-significant and in classical statistics one would conclude that the observed data could plausibly come from a population of mean zero; 'no change'. However, a value of 20 is just as far from the observed mean (10) as is zero. Therefore, the data are just as consistent with the counter-null hypothesis of 20 as they are with the null hypothesis of zero. Nothing in the observed data say that a true population change of 0 is more likely than a change of 20, only the NHST-centric approach gives it this prominence. One can easily imagine a clinical situation in which concluding that the data were consistent with 'no change', when in fact a change of 20 was just as well supported, could be disastrous.

Third, the NHST-centric approach encourages dismissal or acceptance of hypotheses, rather than an assessment of degrees of likelihood. One should ideally design experiments where (the effect size estimates from) the data are likely under one favoured hypothesis but not others. Instead, much biological research sets out to falsify (or,

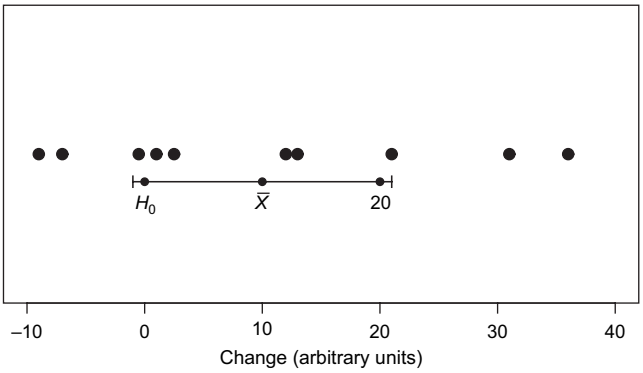

Fig. 1. Illustration of the relationship between the null hypothesis and 'counter-null hypothesis' in a one-sample situation when the null hypothesis (H_0) is zero. When confidence intervals include zero, the null hypothesis is formally not rejected. However the counter-null hypothesis, lying at the same distance on the opposite side of the sample mean ($\bar{X}$), has just as much statistical support as the null hypothesis of zero.

more accurately, render unlikely) the null hypothesis, which is rarely the experimental hypothesis under scrutiny. The danger here is that one ends up 'affirming the consequent', one of the 13 logical fallacies described by Aristotle (Gabbay *et al.*, 2002). A theory, *A*, predicts that a change in *X* causes *Y*; one manipulates *X* and observes *Y* (as supported by a rejection of the null hypothesis); one concludes that theory *A* is supported. This is fallacious, most obviously because theories *B*, *C*, *D* and *E* may also predict that *X* influences *Y* and may even be more likely. Even if the conclusion is the more cautious "our results are consistent with theory *A*", this is weak science. Good science would pit theory *A* against theories *B*, *C*, *D* and *E* with an experiment where each theory gave different predictions. In some areas of biology data are indeed collected with a view to testing plausible alternative hypotheses: within our own discipline of behavioural ecology, sex ratio theory is the prime example (Hardy, 2002) and optimal foraging theory adopted this stance after early criticism (Kacelnik & Cuthill, 1987). However, in too many studies only two hypotheses are aired: the favoured one and the null hypothesis. It is worth highlighting here what the *p* value in NHST represents: the probability of the data (and even more unlikely events) if the null hypothesis is true. Instead, is it not often more interesting to ask what the probability of a given hypothesis is, given the data? The latter, *p*(hypothesis | data) rather than *p*(data | hypothesis), requires a Bayesian approach rather than the classical statistics of NHST (e.g. Yoccoz, 1991; Cohen, 1994; Hilborn & Mangel, 1997).

A likely counter to the arguments in the previous paragraph is that many fields within biology are young disciplines and with new theory one simply wants to know whether there is any effect at all. A *p* value apparently provides the necessary information: the likelihood of getting the observed effects given that the null hypothesis is true [i.e. *p*(data | hypothesis)]. However, with sufficient measurement precision and a large enough sample size one can always obtain a (statistically) non-zero effect. The reason that this jars with the intuition of many biologists is, we feel, the result of multiple meanings of the word 'effect'. Biology, like any science, seeks to establish causal relationships. When biologists talk of an 'effect' they mean a causal influence; they often rely heavily and appropriately on experiments to distinguish cause from correlation. However an effect in statistics need not imply causality; for example, a correlation coefficient is a measure of effect. Measures of the magnitude of an effect in statistics (i.e. effect size; see below) are simply estimates of the differences between groups or the strength of associations between variables. Therefore there is no inconsistency between the statements that a factor has no biological (causal) effect and yet has a measurably non-zero statistical effect.

(2) Effect size and confidence interval

In the literature, the term 'effect size' has several different meanings. Firstly, effect size can mean a statistic which estimates the magnitude of an effect (e.g. mean difference, regression coefficient, Cohen's *d*, correlation coefficient). We refer to this as an 'effect statistic' (it is sometimes called an effect size measurement or index). Secondly, it also means the actual values calculated from certain effect statistics (e.g. mean difference = 30 or $r = 0.7$; in most cases, 'effect size' means this, or is written as 'effect size value'). The third meaning is a relevant interpretation of an estimated magnitude of an effect from the effect statistics. This is sometimes referred to as the biological importance of the effect, or the practical and clinical importance in social and medical sciences.

A confidence interval (CI) is usually interpreted as the range of values that encompass the population or 'true' value, estimated by a certain statistic, with a given probability (e.g. Cohen, 1990; Rice, 1995; Zar, 1999; Quinn & Keough, 2002). For example, if one could replicate the sampling exercise a very large number of times, roughly 95% of the 95% CIs calculated from these samples would be expected to include the true value of the population parameter of interest. The deduction from this being that one can be fairly certain that the value of the population parameter lies within this envelope (with a 5% chance of being wrong, of course). The interpretation that CIs provide an envelope within which the parameter value of interest is likely to lie (e.g. Grafen & Hails, 2002) makes sense even when trying to estimate one-off events for which a 'true' population value has no obvious meaning, such as the probability that a particular species becomes extinct within a given time frame (for Bayesian perspective of CIs or 'credible' intervals, see Clark & Lavine, 2001; Woodworth, 2005; McCarthy, 2007).

The approach of combining point estimation of effect size with CIs provides us with not only information on conventional statistical significance but also information that cannot be obtained from *p* values. For example, when we have a mean difference of 29 with 95% CI = –1 to 59, the result is not statistically significant (at an α level of 0.05) because the CIs include zero, while another mean difference 29 with 95% CI = 9 to 49 is statistically

significant because the CI does not include zero. We stress that the CIs around an effect size are not simply a tool for NHST, but show a range of probable effect size estimates with a given confidence. By contrast, p values allow only a dichotomous decision. While it is true that a dichotomous decision may often be what we need to make in many research contexts, automatic yes-no decisions at $\alpha = 0.05$ can hinder biologists from thinking about and appreciating what their data really mean. As we see later on, consideration of effect size and its CIs will enable researchers to make more biologically relevant decisions.

In addition, researchers should be more interested in how much of an effect their manipulations had and how strong the relationships they observed were than in statistical significance. Effect statistics quantify the size of experimental effects (e.g. mean difference, Cohen's d) and the strength of relationships (e.g. Pearson's r, phi coefficient; see below for more details on effect statistics). Identifying biological importance is what all biologists are ultimately aiming for, not the identification of statistical significance. What is more, dimensionless effect statistics such as d, g, and r (often called standardised effect sizes) set up platforms for comparison among independent studies, which is the basis of meta-analysis.

(3) Encouraging 'meta-analytic' and 'effective' thinking

Since Gene Glass (1976) first introduced meta-analysis, it has become an essential and established tool for literature review and research synthesis in the social and medical sciences (Hunt, 1997; Egger, Smith & Altman, 2001; Hunter & Schmidt, 2004). In evolution and ecology meta-analysis is still fairly new, with meta-analytic reviews starting to appear in the early 90s (e.g. Gurevitch & Hedges, 1993; Arnqvist & Wooster, 1995). Meta-analysis is an effect-size-based review of research that combines results from different studies on the same topic in order to draw general conclusions by estimating the central tendency and variability in effect sizes across these studies. Because of this emphasis, rather than on statistical significance, meta-analysists naturally think outside of the limitations of NHST (Kline, 2004). In social and medical sciences, series of meta-analyses have revealed that the conclusions of some individual studies based on NHST have been wrong (e.g. Lipsey & Wilson, 1993; see also Hunt, 1997). Recently, the benefits of meta-analysis have been described as 'meta-analytic' thinking (Cumming & Finch, 2001; Thompson, 2002*b*). Characteristics of meta-analytic thinking include the following: (1) an accurate understanding of preceding research results in terms of effect size is essential; (2) the report of effect size (along with its CIs) becomes routine, so that results can easily be incorporated into a future meta-analysis; (3) comparisons of new effect sizes with effect sizes from previous studies are made for interpretation of new results, and (4) researchers see their piece of research as a modest contribution to the much larger picture in a research field (for the benefits of Bayesian approach, which somewhat parallels those of meta-analytic thinking, see McCarthy 2007). However, care should be taken with meta-analytic reviews in biology. Biological research can deal with a variety of species in different contexts, whereas in social and medical sciences research is centred around humans and a narrow range of model organisms, often in controlled settings. While meta-analysis of a set of similar experiments on a single species has a clear interpretation, generalization from meta-analysis across species and contexts may be questionable. Nevertheless, meta-analytic thinking itself is a vital practice for biologists.

In meta-analysis, presentation of effect statistics and their CIs is mandatory. Familiarization with effect statistics and their CIs encourages not only meta-analytic thinking but also what we name 'effective' thinking. The benefit of effective thinking is condensed and seen in Fig. 2. As you can see in the figure, the combination of effect sizes and CIs can reveal what p values cannot show (i.e., uncertainty of effect, direction of effect, and magnitude of effect). The approach of using effect sizes and their CIs allows effective statistical inference from data, offering a better understanding and characterisation of the results. It seems that many researchers have fallen for the apparent efficiency of NHST which allows them simple dichotomous decisions (statistically significant or not at $\alpha = 0.05$). It is often the case that a result with $p < 0.05$ is interpreted as representing a real effect whereas a result with a p value larger than 0.05 is interpreted as representing no real effect; this is wrong.

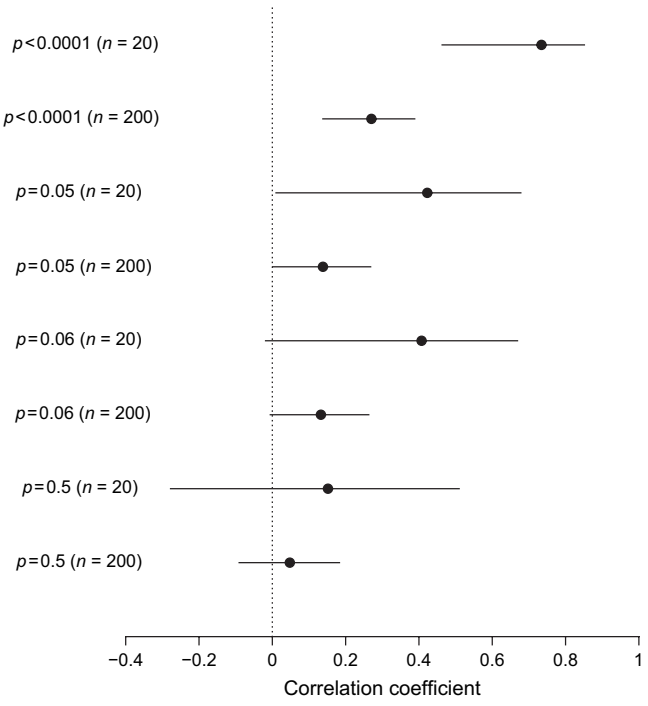

Fig. 2. Effect size estimations (correlation coefficient) and their confidence intervals (CIs). Each pair of p values is based on two different sample sizes: $n = 20$ and $n = 200$. The same p values with different sample sizes can provide dissimilar effect size estimates and their CIs. For example, the two effect size estimations of what is usually termed 'highly significant' p value (i.e. $p > 0.0001$) are remarkably different.

Fig. 2 illustrates that the difference between $p = 0.05$ and $p = 0.06$ in terms of effect size is minimal. What is more, non-rejection of the null hypothesis is frequently interpreted as evidence for no effect without any further evidence for the null hypothesis. Both conclusions are fallacious. When a non-significant result is obtained, the result is only 'inconclusive' (Fisher, 1935; Cohen, 1990). By contrast, the dual approach of including effect sizes and their CIs is effective in interpreting non-significant results. Data analysis that focuses on effect size attributes rather than relying on statistical significance will make biology proceed as cumulative science rather than a series of isolated case studies (for a criticism of the use of adjusted p values, or Bonferroni-type procedures, for multiple comparison, see Nakagawa 2004 and the references therein).

(4) Power analysis is right for the wrong reasons

Effect size is also a crucial component of statistical power analysis. Statistical power analysis utilises the relationships amongst four statistical parameters: sample size, significance criterion (α or the Type I error rate), effect size, and power (which is the probability the test will reject the null hypothesis when the null hypothesis is actually false, or $1 - \beta$, the Type II error rate). When any three of these four parameters are fixed, the remaining one can be determined (Cohen, 1988; Nakagawa & Foster, 2004). Statistical power analysis has gained popularity mainly as a tool for identifying an 'appropriate' sample size. However, power analysis is part of NHST and thus has the associated problems of NHST (e.g. over-emphasis on attainment of statistical significance). Fortunately, power analysis can provide researchers with a good experimental design, albeit for unintended reasons, because the factors which increase power also contribute to an increased precision in estimating effect size (i.e. an increase in sample size generally reduces the CI). Thus power analysis, as part of good experimental design, is right for the wrong reasons (see also Schmidt, 1996; Gelman & Hill, 2007).

III. HOW TO OBTAIN AND INTERPRET EFFECT SIZE

(1) Choice of effect statistics

Kirk (1996) listed more than 40 effect statistics and more recently 61 effect statistics have been identified by Elmore (2001, cited in Huberty, 2002). As effect size reporting becomes obligatory in the social and biomedical sciences, more effect statistics, which are fit for particular sorts of statistical methods, are expected to emerge. For researchers who have never calculated effect size, the task of choosing the appropriate effect statistics for their experimental designs may seem overwhelming. For example, one could go ahead and calculate a single effect statistic for a two-way analysis of variance (ANOVA) with two and five levels in each factor respectively. But how useful will this effect size be in understanding the experimental results? In general, we are ultimately interested in specific relationships (pair-wise group differences or a linear or polynomial trend), not in the combined set of differences among all levels (see Rosenthal *et al.*, 2000). However, we are able to reduce any multiple-level or multiple-variable relationship to a set of two-variable relationships, whatever experimental design we are using (Rosenthal *et al.*, 2000). Therefore, three types of effect statistics suffice for most situations: r statistics (correlation coefficients including Pearson's, Spearman's, point-biserial, and phi; for details, see Rosenthal, 1994; Fern & Monroe, 1996), d statistics (Cohen's d or Hedges' g), and the odds ratio (OR, one of three most used comparative risk measurements, namely odds ratio, relative risk and risk difference; see Fleiss, 1994; Kline, 2004). Calculating and presenting these three effect statistics facilitates future incorporation into a meta-analysis because the methods have been developed to deal especially with these three types of effect statistics (Shadish & Haddock, 1994; Hunt, 1997; Lipsey & Wilson, 2001; Hunter & Schmidt, 2004; note that we will discuss the importance of unstandardised effect statistics below).

The r statistics are usually used when the two variables are continuous; many non-experimental studies are of this type (the distinction between correlation, i.e. r statistics, and regression is discussed below). The d statistics (sometimes referred to as standardised mean differences) are used when the response (dependent) variable is continuous while the predictor (independent variable) is categorical; d should be calculable for pair-wise contrasts within any ANOVA-type design as well as intrinsically two-group studies. The odds ratio is used when the response variable is dichotomous and the predictor variable(s) dichotomous or continuous, such as in contingency tables, logistic regression, loglinear modelling and survival analysis (see Breaugh, 2003; Faraway 2006).

Table 1 lists the most likely cases for d calculations. It is important to notice that d calculations do not change according to whether or not the two groups or treatments are independent, whereas t calculations do. Dunlap *et al.* (1996) point out that many meta-analysists have erroneously used Equation 3 where they should have used Equation 4 (Table 1), inflating effect size unintentionally (see Section III.5 for more on non-independence). Table 2 shows how to obtain the odds ratio and an r statistic for a two by two contingency situation. Odds ratios are also calculated when a predictor variable is continuous. However, this type of odds ratio is not dimensionless (i.e. varies with the units of measurement) and so is less readily comparable across studies. Because an r statistic is calculable in a two by two contingency case, and also to avoid confusion about different applications of odds ratios, we focus only on r and d statistics as standardised measure of effect size in this paper.

However, our focus on these two standardised effect statistics does not mean priority of standardised effect statistics (r or d) over unstandardised effect statistics (regression coefficient or mean difference) and other effect statistics (e.g. odds ratio, relative risk and risk difference). If the original units of measurement are meaningful, the presentation of unstandardised effect statistics is preferable

Table 1. Equations for calculating *d* statistics

Case	Equation		Description	References
Comparing two independent or dependent groups (i.e. both paired and unpaired *t*-test cases)	$d = \frac{m_2 - m_1}{s_{\text{pooled}}}$ $s_{\text{pooled}} = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2}}$	(1) (2)	m_1 and m_2 are means of two groups or treatments, s_{pooled} is pooled standard deviation, n is sample size (in the case of dependent design, the number of data points), s^2 is variance.	Cohen (1988); Hedges (1981)
Comparing two independent groups (i.e. unpaired *t*-test case)	$d = t_{\text{unpaired}}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	(3)	Alternatively, t values can be used to calculate d values; t_{unpaired} is the t value from the unpaired t-test (compare with Equation 10 in the text)	Rosenthal (1994)
Comparing two dependent groups (i.e. paired, or repeated-measure *t*-test case)	$d = t_{\text{paired}}\sqrt{\frac{2(1 - r_{12})}{n}}$	(4)	t_{paired} is the t score from the paired t-test, r_{12} is correlation coefficient between two groups, and note that $n = n_1 = n_2$ not $n = n_1 + n_2$	Dunlap *et al.* (1996)

Free software by David B. Wilson to calculate these effect statistics is downloadable (see Table 4). Strictly speaking, Equations 1 to 4 are for Hedges's *g* but in the literature these formulae are often referred to as *d* or Cohen's *d* while Equation 10 is Cohen's *d* (see Kline, 2004, p.102 for more details; see also Rosenthal, 1994; Cortina & Nouri, 2000).

over that of standardised effect statistics (Wilkinson & the Task Force on Statistical Inference, 1999). For example, imagine we investigated the sex differences in parental care of a species of bird, and found that the difference was $d = 1.0$ with 95% CI = 0.4 to 1.6. It is often more biologically useful to know whether the magnitude of the difference was 1 (95% CI = 0.4 to 1.6), 5 (95% CI = 2 to 8), 10 (95% CI = 4 to 16), or 100 (95% CI = 40 to 160) visits to the nest per hour. If researchers understand their study systems well, original units often help interpretation of effect sizes (see below). Standardised effect statistics are always calculable if sample size and standard deviation are given along with unstandardised effect statistics (see Tables 1 and 2). Also, meta-analysists benefit from knowing the original units, as differences in measured quantities regarding the same subject, say parental care, could results in differences in standardised effect size estimations, which in turn bias the outcome of a meta-analysis (e.g. the use of visits to the nest per hour or amount of food brought to the nest per hour; see Hutton & Williamson, 2000). We would like to point out that, surprisingly, essential pieces of information such as sample sizes and standard deviations are often lacking in research papers and instead there is only the presentation of relevant p values, which themselves are little use for meta-analysis. This problem will be alleviated once researchers appreciate the importance of effect size reporting. There are situations where original scales mean little, or are not readily interpretable, because of a lack of knowledge of the scales or the study systems. In such situations, standardised effect statistics are useful. Choice of standardised or

Table 2. A two by two contingency table for an observed group contrast and equations for calculating odds ratio (OR) and its standard error (*se*) of ln(OR) and an *r* statistic

	Outcome 1	Outcome 2
Group 1	*A*	*B*
Group 2	*C*	*D*

Equation		Description
$p_1 = \frac{A}{A + B}$ (5), $p_2 = \frac{C}{C + D}$	(6)	p_1 and p_2 are a proportion of Outcome 1 in the two groups
$\text{OR} = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{AD}{BC}$	(7)	OR = odds ratio
$se_{\ln(\text{OR})} = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$	(8)	The distribution of OR is not normal but that of ln(OR) is normal.
$r = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} = \sqrt{\frac{\chi_1^2}{n}}$	(9)	sometimes written as φ (phi coefficient), a special case of Pearson's r; $n = A + B + C + D$

The letters *A–D* represent observed cell frequencies. If *A, B, C,* or $D = 0$ in the computation of OR, 0.5 is often added to all cells for correction. Confidence intervals for OR can be calculated using Equations 8 and 15 (see Fleiss, 1994; Rosenthal, 1994; Kline, 2004).

unstandardised effect statistics at this level should be left to researchers, as they are the ones who are the most informed on the relevant measurement scales.

We mainly list calculation procedures for standardised effect statistics (*r* and *d*) below. This is because calculations of standardised effect statistics and their CI are not straightforward compared to their unstandardised counterparts (Smithson, 2001; Thompson, 2002*b*). Most statistical software provides unstandardised effect statistics and their CI so that they do not require special treatment here. General arguments later in our article are applicable to both standardised and unstandardised effect statistics.

There is common confusion and an extremely important point regarding the use of *r* statistics (i.e. correlation) and the distinction from regression. While there are mathematical relationships between some of the formulae used in correlation and regression (see below), their goals and derivations are distinct. Correlation measures association while regression attempts prediction. In the most familiar form of regression analysis, ordinary least squares, two sorts of effect statistics are commonly quoted. The first is coefficient of determination, R^2. It quantifies the proportion or percentage of variation in the response variable that can be accounted for by the predictor(s); it has the advantage that the explanatory power of the independent variable has an immediate intuitive interpretation (e.g. "12% of difference in mating success is attributable to body size" or "23% of variation in maze-learning speed is heritable"). It has the disadvantage that because the magnitude of the R^2 value depends on the original variance 'to be explained', comparison across studies can be misleading (Achen, 1982) or even meaningless (King 1986). So, just because a predictor has a larger R^2 in one situation than another does not mean that the predictor is more influential in the former situation; there may have been less original variation in the first study. Although R^2 appears to be the squared Pearson's correlation coefficient (*r*), and has sometimes been converted to this for meta-analysis, the two are not interchangeable because *r* measures shared variation between *y* and *x*, whereas R^2 is the variation in *y* attributable to (linear variation in) *x*. Taking the square root of R^2 leads to a biased measure of effect (see Equation 13 below).

The second type of effect statistic derived from regression analysis is the slope, *b*, or sometimes standardised slope (termed beta, β, in the widely used statistical package SPSS, although this can lead to confusion because beta is also used to describe the population parameter estimated by the, unstandardised, slope in regression). The slope is the change in the response variable for a unit change in the predictor variable; as the magnitude of *b* depends on the units of measurement it is not a standardised effect statistic. A standardised slope is the change in the response variable, measured in standard deviations, associated with a change of one standard deviation in the predictor variable. It is thus a standardised measure of how much *y* is expected to change when *x* changes by a given amount which, when testing quantitative models of causal influence, is a more natural measure of effect than R^2. However, as argued forcefully by King (1986) and Luskin (1991), if the original goal of a regression analysis is to predict *y* through knowledge of *x*, then why abstract to standardised measures such as R^2 and beta? Therefore we recommend that the unstandardised slope is presented along with its confidence intervals, in addition to R^2 and/or adjusted R^2 (see below; Equation 12). That said, in meta-analysis, a relevant method of incorporating information from a regression analysis may be required and, with careful interpretation and caution, R^2, beta and transformations to *r* can have utility (Luskin, 1991).

(2) Covariates, multiple regression, GLM and effect size calculations

Effect size calculated from two variables is appropriate if there are no influential (or confounding) covariates (e.g. not controlling for effects of sex or weight in a hormonal manipulation experiment; see Garamszegi, 2006). It is possible that even the direction of the effect can change from positive to negative if highly influential covariates exist. In other words, the biological interpretation of effect size statistics can sometimes be completely wrong if we do not consider covariates. In non-experimental studies, many covariates often exist for a predictor variable of interest and, in experimental studies, controlling for a covariate may increase the precision with which one can estimate the experimental effect. Generalized linear models (GLMs; McCullagh & Nelder 1989; Dobson, 2002) provide a common framework for the analysis of models, such as analysis of covariance (ANCOVA), that incorporate both categorical and continuous predictor variables, as well as problems traditionally analysed by ANOVA or regression. As they are an extension of multiple regression, the effect statistics can be derived in an analogous fashion.

Before considering how effect estimates and CIs can be calculated for multiple regression and GLM problems, there is a simple, but absolutely crucial, point to remember. Unless one is analysing a model in which the predictor variables are completely uncorrelated, a condition only likely in a factorial experimental design that is completely balanced, the effect size estimates for a given variable will vary according to what other predictor variables are in the model. For this reason, it could be misleading to compare the slopes, standardised or not, or the (partial) R^2 values for a predictor variable among analyses in which different, or no, other predictor variables are included in the model. This is a specific instance of the general problem, referred to earlier, that estimates of the variance explained by predictor variables depend on the total variance to be explained, and inclusion of additional predictor variables consumes some of that variance.

With the preceding caveat in mind, it is always possible to obtain *t* values from a statistical model for each continuous predictor variable and also for each group (level) of a categorical predictor variable. Generally, *t* values are obtained from a difference between estimates (e.g. means or slopes) divided by the standard error of the differences; almost all statistical software provides *t* values when a statistical model is constructed. The *t* values obtained for groups or categories in a predictor variable can be used for calculating *d* with a formula:

$$d = \frac{t(n_1 + n_2)}{\sqrt{n_1 n_2}\sqrt{df}}, \quad (10)$$

where n_1 and n_2 are the numbers of sample size in two groups and df is the degrees of freedom used for a corresponding t value in a linear model (Equation 10 should be used over Equations 1–3 in Table 1 when t values are obtained from multiple regression; see below). The t values for a continuous predictor variable can be converted to r using a rather unintuitive equation below:

$$r = \frac{t}{\sqrt{t^2 + df}}. \quad (11)$$

Effect size calculated using in this way takes covariates into account. This form of r value is often referred to as a partial correlation coefficient. The partial correlation between y and x_1, controlling for x_2, is numerically equivalent to the correlation between the residuals from the regression of y on x_2 and the residuals from the regression of x_1 on x_2. Thus the partial coefficient for a given predictor removes the variance explained by other predictor variables from both variables, and then quantifies the remaining correlation. A simple case of partial correlation is described below:

$$r_{12|3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}, \quad (12)$$

where $r_{12|3}$ is a partial correlation between variables 1 and 2 controlling for variable 3. As you can imagine, by using Equation 11 (and also Equation 10), we are able to control for a list of covariates. However, the calculation of r from t values introduces bias when predictor variables are non-normal which may often be the case (the bias is analogous to the difference between Pearson's r and Spearman's r when variables are not normal; see Section III.4 dealing with heterogeneous data).

Furthermore, unnecessary predictor variables in the statistical model can influence the estimates of other, perhaps more important, effects. Therefore, careful statistical model selection procedures are essential; in other words, determining what predictors should be in a model and what predictors should be taken out of the model. A problem here is that there seems to be no strong consensus on what is the most appropriate model selection procedure. A popular procedure is to obtain minimum adequate statistical models (based on the principle of Occam's Razor; cf. Whittingham *et al.*, 2006) and there are two common ways of doing so: one using statistical significance (e.g. Crawley, 2002) and the other using the Akaike's information criterion (AIC) (an information-theoretic, IT, approach; Johnson & Omland, 2004; Stephens *et al.*, 2005; note that the IT approach often results in more than one 'important' model, in which parameters, or effect sizes, can be calculated as weighted means according to a weight given to each remaining model; for detailed procedures, see Burnham & Anderson, 2002). An example of the first approach is to achieve model simplification through sequential deletion of the terms in the model that are found to be least statistically significant until all the terms remaining attain statistical significance below some threshold, often $p = 0.1$ (sometimes referred to as the backwards elimination method). The second approach is to find a model which has the smallest AIC value of all models considered. The AIC is an index which weighs the balance between the likelihood of the model and the number of parameters in the model (i.e. a parsimony criterion). The model with the smallest AIC is supposed to retain all influential and important terms, i.e. covariates (as noted above, several competing models with small AIC values, out of all investigated models, are often retained). We should note that the former approach using statistical significance will have the weaknesses of NHST (e.g. influence of sample size). Also the IT approach using AIC is not without problems (see Guthery *et al.*, 2005 for criticisms; see also Stephens *et al.*, 2005; McCarthy 2007). Although both approaches may often result in the same model or similar models, thus providing us with similar effect size estimates, care should be taken in model selection whichever approach is used. Another way of selecting models (and estimating parameters), which is recently gaining popularity in biology, is a Bayesian approach (for more details see Basáñez *et al.*, 2004; Ellison, 2004; Clark, 2005; Clark & Gelfand, 2006; McCarthy, 2007). However, it is worth noting that in more experimental areas of biology the search for a minimum adequate, or the best, model may not be as crucial as in disciplines that are more observational than experimental in nature. When one or more factors are experimentally manipulated the final (and only) model retains these factors to determine their magnitude of effect (see Stephens *et al.*, 2005; Whittingham *et al.*, 2006). Model selection should probably be dictated by the nature of data; biologists should use their experience and expertise to decide what biologically meaningful factors should be in a particular model and, then, see if the direction of the estimated effect of each factor from the model makes sense (see Gelman & Hill, 2007). We will not dwell on model selection any further here since this is not a focus of this paper, but readers are encouraged to explore the literature cited above (see also Faraway, 2005, 2006).

The effect size calculations described above may be extendable to GLMs with binomial, Poisson and other distributions from the exponential family, and with complex error structures (McCullagh & Nelder, 1989; Dobson, 2002). These models usually provide z values instead of t values (i.e. they use the normal distribution rather than the t distribution). We can use obtained z values to replace t values in the relevant equations for calculation of effect size (note that the degrees of freedom should be calculated as if t-tests were used). The use of GLMs is one of several ways which make it possible to calculate effect size from heterogeneous data (i.e. non-normal error structure and/or non-uniform variance; see below for more discussion). However, we are unsure how much bias may be incurred from this procedure in estimating d and r.

We return to a common confusion among researchers regarding R^2, which represents the variance in the data that is accounted for by a particular model. Often, the square-root of R^2 is used as an effect statistic in meta-analysis when models include one predictor, and even when they include

more than one predictor. However, the square-root of R^2 provides a biased effect-size estimate of a predictor of interest and this bias is especially severe when sample size is small. The equation below should be used to correct this bias:

$$r_{\text{adjusted}} = \sqrt{1 - \frac{(n-1)(1-R^2)}{n-k-1}}, \qquad (13)$$

where k is the number of predictors in the model (not including the intercept), n is the sample size (Montgomery & Morrison, 1973); this is the square root of adjusted R^2 which is often calculated in statistical software along with R^2. Although this adjustment may be used for univariate models, it is not desirable to use this effect size estimate for a particular predictor in multivariate models. We recommend effect size be estimated from t values or raw data as suggested above (Table 1). Effect size estimation from models using t values may be used even when there are quadratic or polynomial predictors and interactions. An example of this is considering an interaction between strain type and temperature on the growth of two strains of bacteria. The t value for this interaction between strain and temperature (i.e. difference in slopes) can be used to calculate d using Equation 10 (e.g. $t = 3.1$, $n_1 = n_2 = 30$, $df = 55$ then $d = 0.84$). However, we should be aware that when higher order interactions exist, the main effects (or lower order interactions) of the constituent variables are difficult to interpret in a meaningful way (Crawley, 2002) and thus, the effect size of main effects and lower order interactions requires special care in interpretation. For example, when a model has strain type-by-temperature interaction as an influential factor, the interaction can make effect size estimates for the main effects of strain type and temperature uninterpretable if the slopes for the interaction are merging, diverging or crossing. However, if the slopes are in similar directions, there are cases where effect size estimations from main effects are meaningful. Graphical presentations are often the easiest way to understand the nature of interactions, which is why graphics are given such prominence in statistical software aimed at statisticians (e.g. Venables & Ripley, 2002; Maindonald & Braun, 2003).

(3) Dealing with bias

Two major biases can occur for effect statistics, especially when sample size is small. One is an inherent bias for a particular statistic and the other is a bias caused by sampling errors. The former is of little concern for the correlation coefficient (Hunter & Schmidt, 2004). The d statistics show an upward bias that is relatively large when sample sizes are less than 20 (or less than 10 in each group). Hedges & Olkin (1985) have proposed the equation below to correct this bias.

$$d_{\text{unbiased}} = d_{\text{biased}}\left[1 - \frac{3}{4(n_1+n_2-2)-1}\right], \qquad (14)$$

where n_1 and n_2 are sample sizes of two comparison groups [note that when a paired designed is used, $n_1 = n_2 = n$ so that the denominator can be written as $8(n-1)-1$]; d_{unbiased} is called Hedges' d (d_{biased} is Cohen's d or Hedges' g). It is recommended that this correction be used routinely, although bias is negligible when sample size is large.

The bias incurred by sampling errors is applicable to both r and d statistics and can be severe when sample size is small. CIs, which show the precision of an estimate, are a solution here. Although calculation of CIs for familiar statistics such as means and standard deviations is fairly straightforward, the correct calculation of CIs for effect sizes is not. This is because the construction of CIs around effect size involves the use of non-central t and F distributions, which most biologists have never heard of and for which no generic formulae exist (Thompson, 2002*a*). However, 'traditional' CIs, which offer approximate estimates, are easily calculable. The approximate width of 95% CIs for an effect size is:

$$95\%\text{CI} = ES - 1.96se \text{ to } ES + 1.96se, \qquad (15)$$

where ES stands for effect size (e.g. d, or z-transformed r) and se is the asymptotic standard error for the effect size (note that these formulae are also used for calculations of unstandardised effect statistics and also that t distribution with appropriate df should be used instead of 1.96 when sample size is small, say, less than 20; for simulating CI, see Faraway, 2005, 2006; Gelman & Hill, 2007). The formulae for se are given in Table 3. Fortunately, construction of the exact effect size is easily achievable using computer software (and some programmes calculate the 'exact' CIs around effect sizes). Table 4 lists these programmes and also those that calculate the effect sizes discussed herein. Also, at www.bio.bris.ac.uk/research/behavior/effectsize.htm, we provide scripts written in the free statistical software *R* (www.r-project.org) which include some examples to calculate CIs from simulation and also from bootstrapping which can deal with heterogeneous data.

(4) Problems with heterogeneous data

If data have a heterogeneous (i.e. non-uniform) error structure and variance (e.g. non-parametric data), effect statistics calculated using these data are likely to be biased and the CIs are likely to be inaccurate. Some social scientists have acknowledged this as a major problem with effect size presentation (Grissom & Kim, 2001). There is no consensus on how to deal with this problem although several procedures have been proposed as non-parametric measures of effect size (e.g. Mielke & Berry, 2001; Johnston, Berry & Mielke, 2004; reviewed in Grissom & Kim, 2001).

One obvious solution is the use of transformation and one can calculate standardised effect statistics using these values. In most cases, with appropriate transformation, heterogeneity is curable; normalising transformations, especially Box-Cox transformations, are practical (see Crawley, 2002; Fox, 2002). If normalising transformation fails, a drastic solution may be to calculate effect size using rank-transformed values (Hopkins, 2004). This solution may not be so surprising considering Spearman's rank correlation coefficient uses a similar logic. However, if rank

Table 3. Asymptotic estimates of standard errors (*se*) and other formulae required to calculate confidence intervals

Statistic	Equation		Note	References
d (independent, unpaired)	$se_d = \sqrt{\left(\frac{n_1+n_2-1}{n_1+n_2-3}\right)\left[\left(\frac{4}{n_1+n_2}\right)\left(1+\frac{d^2}{8}\right)\right]}$	(16)	Equation 16 provides *se* for Cohen's *d* while Equation 17 provides *se* for Hedges' *d* (unbiased *d* in Equation 14)	Hunter & Schmidt (2004); Hedges (1981)
	$se_d = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2-2)}}$	(17)		
d (dependent, paired, repeated measure)	$se_d = \sqrt{\frac{2(1-r_{12})}{n} + \frac{d^2}{2(n-1)}}$	(18)	$n = n_1 = n_2$, and r_{12} is correlation coefficient between two groups	Becker (1988)
r (correlation coefficient)	$se_{\mathcal{Z}r} = \frac{1}{\sqrt{n-3}}$	(19)	$\mathcal{Z}r$ is the Fisher transformation of *r* and the distribution of *r* is not normal but that of $\mathcal{Z}r$ is normal	Hedges & Olkin (1985)
	$\mathcal{Z}r = 0.5\ln\left[\frac{(1+r)}{(1-r)}\right]$	(20)		
	$r = \frac{e^{2\mathcal{Z}r}-1}{e^{2\mathcal{Z}r}+1}$	(21)		

Refer to Table 1 for some of the symbols used.

transformation does not alleviate heterogeneity in variances, this method will not estimate correct effect size. Also, if we would like to present unstandardised effect statistics, transformed values should be back-transformed to the original scales (in transformed scales, the interpretation of effect sizes, e.g., regression coefficients or mean differences, is often difficult; for effective interpretation of regression coefficients, see Gelman & Hill, 2007). A related point to make here is transformation in predictors of regression. Although regression models do not assume predictors to be normal, appropriate transformations of predictors can often increase the fit of models (Faraway, 2005, 2006). Also standardised effect statistics will be more accurate with normalised response and predictors if these estimates were to be obtained from *t* values from a regression model, by using Equations 10–11.

We have mentioned above GLMs which can deal with heterogeneous data. The recent growth of GLMs in many biological disciplines may mean that effect size calculation with heterogeneous data should rarely pose problems although, we should repeat, the extent of any bias from this procedure is unknown. Another possible solution is to calculate CIs for effect sizes of heterogeneous data using bootstrapping techniques (yet another solution is the use of Bayesian approaches as CIs or 'credible intervals' can be calculated for parameters which are often difficult or otherwise impossible to estimate; see Gelman & Hill, 2007; MacCarthy 2007 and references therein). Bootstrapping is a computer-intensive re-sampling method. In a bootstrapping procedure, a fixed number of samples are randomly selected from the original data with replacement. When this is repeated many times (e.g. 5000) with a computer, the repeated bootstrap samples produce a distribution of estimates of the statistic of interest and the resulting distribution is used for estimation of CIs (Dixon, 2001; Kline, 2004). It should be noted that small sample size will often give incorrect coverage of CIs. Also, there are several bias-correction methods of calculating CIs in bootstrapping (for a concise summary of methods, see Dixon, 2001). Thus, the interested reader is referred to Davison & Hinkley (1997) and Manly (2007).

A recent, powerful and potentially widely applicable approach based on permutation and randomization is MRPP (multi-response permutation procedures; Mielke & Berry, 2001). The essence of the approach is the distribution of pair-wise distances between data points within groups, or putative groups, and the comparison of these distances with those obtained from all permutations of the data (or approximations thereof). As different distance metrics can be used, no particular distributions are assumed, and the method can be adapted to any of the common statistical models (and several less common ones, such as circular statistics and tests of sequential dependence; Mielke & Berry, 2001), the approach has huge potential (for an application in biology, see Endler & Mielke, 2005). The effect statistics that are generated can often be related to the more familiar types already discussed, but only when analogous distance metrics are used (e.g. squared Euclidean distance, the minimisation of which is the model-fitting criterion for most of the classical statistical tests discussed). This is an area for further research; free Windows implementation of MRPP and other related approaches is available at www.fort.usgs.gov/Products/Software/Blossom/.

(5) Non-independence of data

Literature on effect size estimation with non-independent data seems scarce (Dunlap *et al.*, 1996; Hunter & Schmidt, 2004). We have already mentioned the case where control and experimental groups are related (i.e. repeated, matched or correlated designs; see Table 1). There are many other cases in which data points are not independent and some of these may be highly complex (e.g. hierarchical nested or crossed data structures). Here we mention two cases in which we can reasonably estimate effect size. The first case is comparing two groups in which all data points are related in each group. For example, we want to compare decline in sperm velocity over time in two species of birds. The velocities of sperm from 10 individuals of each species are measured every hour for 20 h. Assuming, for argument's sake, that sperm velocity shows a linear decline with time, a regression slope can be calculated for each

Table 4. Lists of selected programs (or websites) which calculate effect statistics and their confidence intervals

Program	Description	Free?	URL address	Source
es calculator	Calculations for *d*, *r*, and other effect statistics	Yes	http://mason.gmu.edu/~dwilsonb/ma.html	David B. Wilson
Effect Size Calculator	Calculations for *d* and its approximate CIs	Yes	http://davidmlane.com/hyperstat/effect_size.htm	Robert Coe
ES	Calculations for *d*, *r*, and other effect statistics	No (free trial version)	http://www.assess.com/Software/ES.htm	Shadish *et al.* (1999)
Smithson's *SPSS*, *SAS*, *S-Plus* and *R* macros	Calculations for exact CIs for *d*, *r* (*via t* and *F* values) and others	Yes	http://www.anu.edu.au/psychology/staff/mike/CIstuff/CI.html	Smithson (2001)
ESCI	Calculations for exact CIs for *d* and *r* (*via t* and *F* values)	No (free trial version)	http://www.latrobe.edu.au/psy/esci/index.html	Cumming & Finch (2001)
STATISTICA (Power Analysis module)	Calculations for exact CIs for *d*, *r*, and others (*via t* and *F* values)	No	http://www.statsoftinc.com/	StatSoft

R scripts including most of effect statistics and CIs used in this article are obtainable at www.bio.bris.ac.uk/research/behavior/effectsize.htm. *R* packages such as 'psychometric' (Fletcher, 2007) include useful functions to calculate standardised effect statistics and CIs. Also, software for meta-analysis generally calculates many types of effect statistics (e.g. *MetaWin* by Rosenberg *et al.*, 2000).

individual. The slopes of the 10 regression lines on time for each of these two species can be compared using a *t*-test. Then, Equation 3 can be used to calculate *d* statistics.

The second case is more difficult to deal with. This is when two groups have repeated measurements and/or related data (two groups include data points from a particular individual and/or each group includes more than one data point from a particular individual). For example, in an ecological field experiment, we want to evaluate the annual breeding success of individuals (the number of offspring produced in each year) when exposed to two experimental treatments. Because of logistical constraints in the field and limited availability of study animals, the experiment takes place over three breeding seasons, with some individuals appearing in both experimental and control treatments (in different years) while others are only in either one of the two treatment groups. Mixed-effects models, which can incorporate information about non-independence as well as covariates, are often used to deal with this type of data (Pinheiro & Bates, 2000; see also Paterson & Lello, 2003). The *t* values from mixed-effects models can be used to approximate *d* statistics using the equations below:

$$d = \frac{t_{\mathrm{MEM}}\left[1 + \left(n_{\mathrm{i}}/n_{\mathrm{o}}\right)R\right]\sqrt{1 - R}\left(n_{\mathrm{o1}} + n_{\mathrm{o2}}\right)}{\sqrt{n_{\mathrm{o1}}n_{\mathrm{o2}}}\sqrt{n_{\mathrm{o}} - k}}, \qquad (22)$$

$$R = \frac{s^2_{\mathrm{B}}}{s^2_{\mathrm{B}} + s^2_{\mathrm{E}}}, \qquad (23)$$

where t_{MEM} is *t* value from mixed-effects model, n_{o1} and n_{o2} are the numbers of observations in each treatment (i.e. one individual may contribute more than one observation), n_{o} and n_{i} are the total number of observations and the number of individuals (or groups), respectively ($n_{\mathrm{o1}} + n_{\mathrm{o2}} = n_{\mathrm{o}}$), *k* is the number of parameters (including the intercept) and *R* is often called the repeatability or intra-class correlation (e.g. Zar, 1999) which consists of two variance components: s^2_{B} (between-individual, or between-group, variance) and s^2_{E} (within-individual, or within-group, variance or residual variance; they are obtained from the random-effect part of the mixed-effects model). In a similar manner, when a predictor variable of interest is continuous (e.g. the effect of temperature on lizard behaviour, where data consist of 10 replicates from each of five animals), the *t* values from mixed-effects models can be used to approximate *r* statistics:

$$r = \frac{t_{\mathrm{MEM}}\left[1 + \left(n_{\mathrm{i}}/n_{\mathrm{o}}\right)R\right]\sqrt{1 - R}}{\sqrt{t^2_{\mathrm{MEM}}\left[1 + \left(n_{\mathrm{i}}/n_{\mathrm{o}}\right)R\right]^2\left(1 - R\right) + n_{\mathrm{o}} - k}}. \qquad (24)$$

CIs for these effect statistics need to be computed using the programmes listed in Table 4 through *t* values from mixed-effects models and their *df*. The procedures of effect size estimation proposed above using mixed-models may or may not work depending on the nature of the data (e.g. structure of the pseudo-replication and sample size). Also,

the way mixed-effects models calculate *df* is different from conventional ANOVAs (sometimes it differs among software) so that CI estimation may not be reliable. Alternatively, approximate CIs for these effect statistics can be calculated from converting CIs for unstandardised measurements, which are usually calculated automatically in statistical software (but this has the same problem as these CIs depends on how software calculates *df*; for a solution using simulation or bootstrapping, see Faraway, 2006; for a Bayesian solution, see Gelman & Hill, 2007). A simpler approach to estimate standardised effect statistics (or a better point estimate) is to use linear model or GLM frameworks. As mentioned before, *d* calculations do not depend on independence of data, but *t* calculations do (see Equations 1–4 Table 1; note that CI calculations for *d* do depend on independence of data, Equations 16-18 in Table 3). Thus, fitting linear models or GLMs to a particular set of data with a certain non-independent structure (by ignoring non-independence) provide *t* values of interest, which can be converted to *d* or *r* point estimates (Equations 10–11). However, this straightforward method does not provide correct CIs (or more simply *d* and *r* can be calculated from raw data, although using raw data does not control for covariates). Although point estimates from this can be used to compare those obtained by using mixed-effects models and their CIs estimates above, this is obviously not a solution to the problem (note that Equations 22–24 only provide approximates; also, for bootstrap for non-independent data, see Davison & Hinkley, 1997).

Incidentally, by extending what is described about GLMs above, it may be possible to calculate standardised effect statistics and their approximate CIs for some generalized linear mixed models (GLMMs), which are increasingly used in biology (note that s^2_{E} in Equation 23 can be set to be 1 for probit-link or $\pi^2/3$ for logit-link GLMMs with binomial errors; Snijders & Bosker, 1999). However, when a mixed-effects model framework is used, currently, it is probably much easier to present unstandardised effect statistics and interpret them. Kline (2004) states that methods for calculating (standardised) effect statistics and its CIs for complicated non-independent designs are lacking. However, we are confident that answers for this particular problem will be forthcoming.

(6) Translating effect size into biological importance

There is little point presenting effect sizes in papers if these are not interpreted and discussed. Thus it is important to know what magnitude of effect size constitutes something biologically important. If researchers are familiar with their study systems, or abundant previous research on a topic of interest exists, effect sizes in original units are more readily interpretable than standardised effect statistics. Comparisons of effect size values between previous research and current work is often fruitful if effect size estimations are in the same units. However, prior knowledge is not always available, or one's research may use different measurement scales from previous research. In such cases, interpreting standardised effect sizes may make more sense as there are some guidelines we can follow. Cohen (1988) has proposed 'conventional' values as benchmarks for what are considered to be 'small', 'medium', and 'large' effects ($r = 0.1$, 0.3, 0.5 and $d = 0.2$, 0.5, 0.8, respectively). However, these benchmarks have been criticized in the social and medical sciences because practical and clinical importance depends on the situation researchers are dealing with (Thompson, 2002*a*, *b*; Kline, 2004). For example, the relationship between cigarette smoking and lung cancer ($r = 0.1$) is considered practically and clinically very important because appropriate legal policy change might save millions of lives (Gage, 1978). By contrast, the same degree of relationship between cigarette smoking and sleeping hours would not be considered practically or clinically very important; it is hard to imagine that a ban on smoking would happen on the basis of this finding.

In terms of pure science, however, these two findings are both interesting and may be considered important as long as both results have narrow and similar CIs. We argue that biological importance is more objective than practical or clinical importance in which subjective (and sometimes political) judgements may be inevitable. Although we have no intention of advocating total reliance on benchmark values in biology, we suggest that benchmarks for effect statistics may be useful. Nonetheless, biologists also should take caution in using benchmarks and should evaluate their effect sizes in the light of their hypotheses and also of results from previous relevant studies. We emphasize the point made by Thompson (2001) who stated that if we use these fixed benchmarks with the same rigidity that $\alpha = 0.05$ is used in NHST, we are just being stupid in another metric.

In this paper, we have emphasised the dual approach of using effect size and its CI. Interpreting the point estimate of effect size itself, without consideration of its CI, may not make sense at all. If a large effect, say $d = 1.2$, has a large CI (95% = 0.1 to 2.3) and another similarly large effect has a small CI (95% = 1.0 to 1.4), the interpretation of these putatively large effects will be different. We think visual presentation of effect size values and their CIs is a useful approach as described in Fig. 2. This visual approach is particularly useful for pair-wise contrasts, i.e. standardised and unstandardised mean difference, in experimental studies. Providing the precision of effect (CI) is essential although it has attracted less attention than the point estimate (effect size) in general (e.g. Wilkinson & the Task Force on Statistical Inference, 1999).

Some people have suggested converting *d* to *r* when *d* statistics are interpreted, because many researchers have some degree of conceptual understanding of *r* statistics and may find it easy to interpret effect size in *r*-converted form (e.g., Cortina & Nouri, 2000; Jennions & Møller, 2003). We recommend effect size estimates be interpreted in their original form because conversions may unnecessarily incur bias (Thompson, 2002*b*) and also it makes more sense to interpret, say, a difference between two groups as *d* rather than *r*. However, we agree that conversions are conceptually helpful and also an essential technique for meta-analysis when integrating the results of studies which have employed different methods (e.g. a correlational approach and

a two-group design). Conversion formulae are below (Rosenthal, 1994):

$$r = \frac{d}{\sqrt{d^2 + \frac{(n_1 + n_2)^2}{n_1 n_2}}}, \quad (25)$$

$$d = \frac{2r}{\sqrt{1 - r^2}}. \quad (26)$$

We should note that correct conversion formulae for Hedges' g are somewhat different from those described above (d is here Cohen's d). The interested reader is referred to Rosenthal (1994) and Fleiss (1994).

IV. CONCLUSIONS

(1) The presentation of effect size along with its CI is urgently required because effect size and its CI provide the two most important pieces of statistical information for biologists: the magnitude estimate of an effect of interest and the precision of that estimate. There is no doubt that the presentation and interpretation of effect size will reduce prevalent misuse and misinterpretation of NHST and the p value in biology. Effect size presentation along with its CI will also benefit and advance our fields as cumulative science, encouraging 'effective' as well as 'meta-analytic' thinking, as is already happening in some other disciplines. The dual approach of presenting both effect size and its CI is essential although the presentation of the CI is less discussed.

(2) Although this article covers many situations for effect size calculation and deals with the problems associated with effect size and its CI calculation and presentation (e.g. the existence of covariates, bias in calculation, non-normality in data, non-independence of data), our article by no means provides comprehensive guidelines. This is a broad topic comprising many issues (see Fern & Monroe, 1996).

(3) Our article, however, serves as a beginner's manual and a starting point for changing statistical practice in biology for the better. In the future, as more and more people report effect sizes, the problems which we could not provide definitive solutions to here will hopefully be solved (and hopefully, effect size and its associated calculations will be more prevalent in common statistical software). Also, as we focus on the calculation of standardised effect statistics that are the basis for meta-analysis, our article serves as a reference when conducting such analyses.

(4) Our particular focus on the two classes of standardised effect statistics (r and d) in this article does not necessarily represent our view of which effect statistic is considered the most important; as we have seen, in some cases, calculations of the standardised effect statistics are complicated. Unstandardised effect statistics (regression coefficient or mean difference) and other effect statistics (e.g. odds ratio) should also be used and presented accordingly. The rule of thumb may be the usage of an effect statistic, which can be interpreted in a biologically meaningful way, depending on biological systems or questions researchers are dealing with. This also relates to the difficulty of biological interpretation of effect size, which is often context-dependent.

(5) Emergent alternative approaches to NHST such as the information-theoretic, IT, and Bayesian approaches may replace NHST in many areas in the future (for more on these alternatives, see e.g. Johnson & Omland, 2004; Ellison, 2004; McCarthy 2007). Whatever inferential statistical approach is used in the future, effect size estimation is here to stay because effect size is the information that all scientists should be interested in, because it relates to biological importance. We repeat that the obligatory presentation of effect sizes with CIs is strongly recommended in any journal of biology. Editors of journals should accept the fact that such presentations may require more space per article, but this is for the betterment of their fields.

V. ACKNOWLEDGEMENTS

Our special thanks to Will Hopkins – without his continuous and enormous inputs and help, this article would not be possible. Comments from statisticians and biologists alike, notably Jane Hutton, Michael Festing, Clare Stanford, László Garamszegi, Simone Immler, Jarrod Hadfield, Phil Stephens, Michelle Simeoni, Ben Hatchwell and his discussion group members, and two anonymous referees, have greatly improved the manuscript. S.N. is supported by the Tertiary Education Commission, New Zealand.

VI. REFERENCES

Achen, C. (1982). *Interpreting and Using Regression*. Sage, Beverly Hills, CA.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P. C. & Lang, T. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* **134**, 663–694.

American Psychological Association (2001). *Publication Manual of the American Psychological Association*, 5th edition. Author, Washington, DC.

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

Basáñez, M. G., Marshall, C., Carabin, H., Gyorkos, T. & Joseph, L. (2004). Bayesian statistics for parasitologists. *Trends in Parasitology* **20**, 85–91.

Becker, B. J. (1988). Synthesizing standardized mean change measures. *British Journal of Mathematical and Statistical Psychology* **41**, 257–278.

Breaugh, J. A. (2003). Effect size estimation: factors to consider and mistakes to avoid. *Journal of Management* **29**, 79–97.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd edition. Springer-Verlag, Berlin.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters* **8**, 2–14.

Clark, J. S. & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology and Evolution* **21**, 375–380.

Clark, J. S. & Lavine, M. (2001). Bayesian Statistics: Estimating Plant Demographic Parameters. In *Design and Analysis of Ecological Experiments* (eds. S. M. Scheiner and J. Gurevitch), pp. 327–346. Oxford University Press, Oxford.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Erlbaum, Hillsdale, NJ.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist* **45**, 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist* **49**, 997–1003.

Cortina, J. M. & Nouri, H. (2000). *Effect Size for ANOVA Designs*. Sage, Thousand Oaks, CA.

Crawley, M. J. (2002). *Statistical Computing: an Introduction to Data Analysis Using S-Plus*. Wiley, Chichester.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement* **61**, 532–574.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

Dixon, P. M. (2001). The bootstrap and the jackknife: describing the precision of ecological indices. In *Design and Analysis of Ecological Experiments* (eds. S. M. Scheiner and J. Gurevitch), pp. 267-288. Oxford University Press, Oxford.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd edition. CRC, Boca Raton, FL.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B. & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measured designs. *Psychological Methods* **1**, 170–177.

Egger, M., Smith, G. D. & Altman, D. G. (2001). *Systematic Reviews in Health Care: Meta-Analysis in Context*. BMJ, London.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters* **7**, 509–520.

Endler, J. A. & Mielke, P. W. (2005). Comparing entire colour patterns as birds see them. *Biological Journal of the Linnean Society* **86**, 405–431.

Faraway, J. J. (2005). *Linear Models with R*. Chapman & Hall/CRC, Boca Raton, FL.

Faraway, J. J. (2006). *Extending the Linear Model with R*. CRC, Boca Raton, FL.

Fern, E. F. & Monroe, K. B. (1996). Effect-size estimates: issues and problems in interpretations. *Journal of Consumer Research* **23**, 89–105.

Fidler, F., Burgman, M., Cumming, G., Buttrose, R. & Thomason, N. (2006). Impact of criticism of null hypothesis significance testing on statistical reporting practices in coservation biology. *Conservation Biology* **20**, 1539–1544.

Fidler, F., Cumming, G., Burgman, M. & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics* **33**, 615–630.

Fisher, R. A. (1935). *The Design of Experiments*. Hafner, New York, NY.

Fletcher, T. D. (2007). *The psychometric Package: Appliced Psychometric Theory*, R package, version 0.1.2. cran.r-project.org/doc/packages/psychometric/.pdf.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In *The Handbook of Research Synthesis* (eds. H. Cooper and L. V. Hedges), pp. 245–260. Sage, New York, NY.

Fox, J. (2002). *An R and S-Plus Companion to Applied Regression*. Sage, Thousand Oaks, CA.

Gabbay, D. M., Johnson, R. H., Ohlback, H. J. & Woods, J. (2002). *Handbook of the logic of argument and inference*. North Holland, Amsterdam.

Gage, N. L. (1978). *The Scientific Basis of the Art of Teaching*. Teachers College Press, New York, NY.

Garamszegi, L. Z. (2006). Comparing effect sizes across variables: generalization without the need for Bonferroni correction. *Behavioral Ecology* **17**, 682–687.

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge

Glass, G. V. (1976). Integrating findings: the meta-analysis of research. In *Review of Research in Education Vol. 5* (ed. L. Shulman), pp. 351–379. Peacock, Itasca, IL.

Grafen, A. & Hails, R. (2002). *Modern Statistics for the Life Sciences*. Oxford University Press, Oxford.

Grissom, R. J. & Kim, J. J. (2001). Review of assumptions of problems in the appropriate conceptualization of effect size. *Psychological Methods* **6**, 135–146.

Grissom, R. J. & Kim, J. J. (2005). *Effect Sizes for Research: a Broad Practical Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

Gurevitch, J. & Hedges, L. V. (1993). Meta-analysis: combining the results of independent experiments. In *Design and Analysis of Ecological Experiments* (eds. S. M. Scheiner and J. Gurevitch), pp. 347–369. Chapman & Hall, New York, NY.

Guthery, F. S., Brennan, L. A., Peterson, M. J. & Lusk, J. J. (2005). Information theory in wildlife science: critique and viewpoint. *Journal of Wildlife Management* **69**, 457–465.

Hardy, I. C. W. (2002). *Sex Ratios: Concepts and Research Methods*. Cambridge University Press, Cambridge.

Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (1997). *What If There Were No Significance Tests?* Erlbaum, Mahwah, New Jersey.

Hedges, L. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York, NY.

Hedges, L. V. (1981). Distributional theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* **6**, 107–128.

Hilborn, R. & Mangel, M. (1997). *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, NJ.

Hopkins, W. G. (2004). *New View of Statistics*. http://www.sportsci.org/resource/stats/.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement* **62**, 227–240.

Hunt, M. (1997). *How Science Takes Stock: the Story of Meta-Analysis*. Russell Sage, New York, NY.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Finding*, 2nd edition. Sage, Thousand Oaks, CA.

Hutton, J. L. & Williamson, P. R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society, Series C, Applied Statistics* **49**, 359–370.

Jennions, M. D. & Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* **14**, 434–445.

Johnson, J. B. & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution* **19**, 101–108.

Johnston, J. E., Berry, K. J. & Mielke, P. W. (2004). A measure of effect size for experimental designs with heterogeneous variances. *Perceptual and Motor Skills* **98**, 3–18.

KACELNIK, A. & CUTHILL, I. C. (1987). Starlings and optimal foraging theory: modelling in a fractal world. In *Foraging Behavior* (eds. A. Kamil, J. R. Krebs and H. Pulliamm), pp. 303–333. Plenum, New York, NY.

KING, G. (1986). How not to lie with statistics: avoiding common mistakes in quantitative political science. *American Journal of Political Science* **30**, 666–687.

KIRK, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement* **56**, 746–759.

KLINE, R. B. (2004). *Beyond Significance Testing*. American Psychological Association, Washington, DC.

LIPSEY, M. W. & WILSON, D. B. (1993). The efficacy of psychological educational, and behavioral treatment: conformation from meta-analysis. *American Psychologist* **48**, 1181–1209.

LIPSEY, M. W. & WILSON, D. B. (2001). *Practical Meta-Analysis*. Sage, Beverly Hills, CA.

LUSKIN, R. C. (1991). Abusus non tollit usum: standardized coefficients, correlations, and R^2s. *American Journal of Political Science* **35**, 1032–1046.

MAINDONALD, J. & BRAUN, J. (2003). *Data Analysis and Graphics Using R: an Example-Based Approach*. Cambridge University Press, Cambridge.

MANLY, B. R. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edition. Chapman & Hall/CRC, Boca Raton, FL.

MCCARTHY, M. A. (2007). *Bayesian Methods for Ecology*. Cambridge University Press, Cambridge.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall, New York, NY.

MIELKE, P. W. & BERRY, K. J. (2001). *Permutation Methods: A Distance Function Approach*. Springer-Verlag, New York, NY.

MONTGOMERY, D. B. & MORRISON, D. G. (1973). A note on adjusting R^2. *Journal of Finance* **28**, 1009–1013.

NAKAGAWA, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology* **15**, 1044–1045.

NAKAGAWA, S. & FOSTER, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica* **7**, 103–108.

NICKERSON, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* **5**, 241–301.

PATERSON, S. & LELLO, J. (2003). Mixed models: getting the best use of parasitological data. *Trends in Parasitology* **19**, 370–375.

PINHEIRO, J. C. & BATES, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York, NY.

QUINN, G. P. & KEOUGH, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge, Cambridge University Press.

RICE, J. A. (1995). *Mathematical Statistics and Data Analysis*. 2nd edition. Duxbury Press, Belmont, CA.

ROSENBERG, M. S., ADAMS, D. C. & GUREVITCH, J. (2000). *MetaWin: Statistical Software for Meta-Analysis*. Sinauer, Sunderland, MA.

ROSENTHAL, R. (1994). Parametric measures of effect size. In *The Handbook of Research Synthesis* (eds. H. Cooper and L. V. Hedges), pp. 231–244. Sage, New York, NY.

ROSENTHAL, R., ROSNOW, R. & RUBIN, D. B. (2000). *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge University Press, Cambridge.

SCHMIDT, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* **1**, 115–129.

SHADISH, W. R. & HADDOCK, C. K. (1994). Combining estimates of effect size. In *Handbook of Research Synthesis* (eds. H. Cooper and L. V. Hedges), pp. 261–282. Russell Sage, New York, NY.

SHADISH, W. R., ROBINSON, L. & LU, C. (1999). *ES*. Assessment Systems, St. Paul, MN.

SMITHSON, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: the importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement* **61**, 605–632.

SNIJDERS, T. & BOSKER, R. (1999). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. Sage, London.

STEPHENS, P. A., BUSKIRK, S. W. & DEL RIO, C. M. (2007). Inference in ecology and evolution. *Trends in Ecology & Evolution* **22**, 192–197.

STEPHENS, P. A., BUSKIRK, S. W., HAYWARD, G. D. & DEL RIO, C. M. (2005). Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology* **42**, 4–12.

THOMPSON, B. (2001). Significance, effect sizes, stepwise methods, and other issues: strong arguments move the field. *Journal of Experimental Education* **70**, 80–93.

THOMPSON, B. (2002*a*). "Statistical", "practical", and "clinical": how many kinds of significance do couselors need to consider. *Journal of Counseling & Development* **80**, 64–71.

THOMPSON, B. (2002*b*). What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher* **31**, 25–32.

VENABLES, W. N. & RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th edition. Springer, New York, NY.

WILKINSON, L & THE TASK FORCE ON STATISTICAL INFERENCE. (1999). Statistical methods in psychology journals. *American Psychologist* **54**, 594–604.

WHITTINGHAM, M. J., STEPHENS, P. A., BRADBURY, R. B. & FRECKLETON, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* **75**, 1182–1189.

WOODWORTH, G. G. (2004). *Biostatistics: a Bayesian Introduction*. Wiley, Hoboken, NJ.

YOCCOZ, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* **72**, 106–111.

ZAR, J. (1999). *Biostatistical Analysis*, 4th edition. Prentice-Hall, Upper Saddle River, NJ.