Behavioral Ecology
doi:10.1093/beheco/ark005
Advance Access publication 28 April 2006

Forum

Comparing effect sizes across variables: generalization without the need for Bonferroni correction

László Zsolt Garamszegi
Department of Biology, University of Antwerp, Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

Studies in behavioral ecology often investigate several traits and then apply multiple statistical tests to discover their pairwise associations. Traditionally, such approaches require the adjustment of individual significance levels because as more statistical tests are performed the greater the likelihood that Type I errors are committed (i.e., rejecting H_0 when it is true) (Rice 1989). Bonferroni correction that lowers the critical *P* values for each particular test based on the number of tests to be performed is frequently used to reduce problems associated with multiple comparisons (Cabin and Mitchell 2000). However, this procedure dramatically increases the risk of committing Type II errors as it results in a high risk of not rejecting a H_0 when it is false. To reach 80% statistical power, it is necessary to have huge sample sizes to detect medium ($r = 0.3$ or $d = 0.5$; sensu Cohen 1988) or small ($r = 0.1$ or $d = 0.2$; sensu Cohen 1988) strength effects (e.g., say $N = 128$ or $N = 788$, respectively, for a 2-sample *t*-test), but sample size is often limited when studying behavior.

The strict application of Bonferroni correction in the field of ecology and behavioral ecology has therefore been criticized for mathematical and logical reasons (Wright 1992; Benjamini and Hochberg 1995; Perneger 1998; Moran 2003; Nakagawa 2004). As a potential solution, Wright (1992) and Chandler (1995) advocated that the sacrificial loss of power can be avoided by choosing an experimentwise error rate higher than the usually accepted 5%, which results in a balance between different types of errors. As another alternative, the researcher might be more interested in controlling the proportion of erroneously rejected null hypotheses, the so-called false discovery rate, than in controlling for familywise error rate (Benjamini and Hochberg, 1995). Although this approach allows for increased power in large series of repeated tests, it is rarely applied in ecological studies (Garcia 2003, 2004).

Recently, Nakagawa (2004) suggested reporting effect sizes together with confidence intervals (CIs) for all potential relationships to allow the readers to judge the biological importance of the results and to reduce publication bias. Due to the low power of the tests, the majority of investigated relationships are expected to be nonsignificant, which is thought to make publication difficult. Such difficulty is generally assumed to cause behavioral ecologists to selectively report data (Moran 2003; Nakagawa 2004). The omission of nonsignificant results from publications is undesirable for both scientific and ethical reasons, which makes Bonferroni adjustment problematic. It is noteworthy that direct tests comparing effect sizes of representative samples of published and unpublished studies showed no evidence of publication bias in the biological literature (Koricheva 2003; Møller et al. 2005). However, independent of publication bias, conclusions drawn from effect sizes and the associated CIs should be encouraged. Such an approach considers the magnitude of an effect on a continuous scale, whereas conventional hypothesis testing based on significance levels tends to treat biological questions as all-or-nothing effects depending on whether *P* values exceed the critical limit or not (Chow 1988; Wilkinson and Task Force Stat Inference 1999; Thompson 2002). Hence, using the same data, the former approach may reveal that a particular effect is small, but still biologically important, whereas, the later approach may lead the investigator to conclude that the hypothesized phenomenon does not exist in nature. Although such philosophical differences may dramatically influence our knowledge, presenting standardized effect sizes is still uncommon in ecology and evolution (Nakagawa 2004).

Here, I suggest that, in addition to their presentation, the calculated effect sizes may be further used in simple analyses that can help to estimate the true effect of a predictor variable and thus make general conclusions. These analytical tools rely on the fact that the strength and direction of relationships, as reflected by standardized measures of effect sizes (Pearson's *r*, Cohen's *d*, or Hedges' *g*), are comparable and independent of the scale on which the variables were measured (e.g., Hedges and Olkin 1985; Cohen 1988; Rosenthal 1991). Thus, if multiple traits are measured and multiple correlations are calculated, the corresponding effect sizes tabulated among the variables measured will have a certain statistical distribution with measurable attributes. Below, I present 4 simple analyses to demonstrate how such statistical attributes can be used to make general interpretations. I will confine myself to a typical sampling design from behavioral ecology in which the experimenter is interested in explaining variation in certain traits (response variables) in the light of other (predictor) variables. Specific sampling designs can be tailored according to the biological question at hand that will be illustrated by using real data on the collared flycatcher, *Ficedula albicollis* from Garamszegi et al. (2004). I will also discuss the confounding effect of colinearity between variables that may violate the assumption of statistical independence and the potentially low power of the suggested tests.

ANALYSES OF EFFECT SIZES

First, the mean effect size from multiple pairwise tests can be calculated to test the null hypothesis that the mean underlying effect size does not differ from zero. It will be rejected if the measured variables covary with a predictor variable consistently in the same direction. Normally, a few of the investigated relationships will be significant but the majority will not (see an example in Table 1). The classical interpretation of these results relies on the relationships that pass the filter of Bonferroni correction (i.e., strong effects). However, weak effects may also have biological importance: a meta-analysis of meta-analyses in ecology and evolution revealed small to intermediate mean effect sizes ($r < 0.2$) and that the amount of variance explained in biological studies appears to be very small (Møller and Jennions 2002). Therefore, neglecting small effects could be misleading as it may cause us to overlook weak but evolutionarily still important patterns. A consistent pattern of variation in all measured effect sizes in a certain

Table 1

Effect sizes (Hedges' *g*) and the associated 95% CIs for the relationship between male sexual traits and success in nest-box retention reflecting success in male–male competition in the collared flycatcher (data from Garamszegi et al. 2004, see methodological details therein)

	Nest-box retention				
Variable (mean ± SD)	Bias-corrected effect size (Hedges' *g*)	*N*	*P*	CI lower	CI upper
Full repertoire size (50.64 ± 23.48)	0.317	26	0.413	−0.459	1.093
Song rate (1/min) (4.06 ± 1.60)	0.834	27	0.032	0.043	1.625
Versatility (%) (0.72 ± 0.07)	0.148	35	0.673	−0.551	0.847
Strophe tempo (1/s) (3.17 ± 0.27)	−0.092	35	0.793	−0.790	0.606
Strophe length (s) (3.34 ± 0.59)	0.049	35	0.889	−0.649	0.747
Strophe repertoire size (7.40 ± 1.34)	0.100	35	0.776	−0.599	0.798
Frequency range (kHz) (6.40 ± 0.47)	0.407	35	0.251	−0.298	1.111
Frequency maximum (kHz) (8.57 ± 0.35)	0.079	35	0.823	−0.620	0.777
Frequency minimum[a] (kHz) (2.17 ± 0.27)	0.623	35	0.086	−0.090	1.336
No. of figures (10.68 ± 2.24)	0.011	35	0.976	−0.687	0.709
Forehead patch size (mm^2) (67.05 ± 15.26)	0.056	35	0.873	−0.642	0.754
Wing patch size[b] (mm) (451.6 ± 151.1)	−0.049	26	0.917	−1.015	0.917

SD: standard deviation. Among the 12 sexual traits, only song rate varied significantly with nest-box retention success. This association is not significant after the traditional Bonferroni correction ($P > 0.0041$). Hence, from the current data, one may conclude that there is no relationship between sexual signaling and nest-box retention success. However, the mean effect size was significantly larger than zero (mean ± SD = 0.208 ± 0.283, $t_{11} = 2.537$, $P = 0.027$) indicating a directional trend. This was also the tendency when I excluded the significant effect for song rate (mean ± SD = 0.153 ± 0.220, $t_{10} = 2.309$, $P = 0.044$), suggesting that individuals with elaborate sexual traits generally have increased success in male–male competition. In these analyses, for each variable I assumed that larger values reflect higher elaborateness. In the majority of cases (e.g., repertoire size, strophe length, song rate, and white patches on the plumage), it is reasonable to assume that large values reflect high male quality. However, for other variables (e.g., frequency minimum and tempo), the biological relevance of low and large values is less obvious. The mean of unsigned effect sizes, which are not confounded by directional definitions, was 0.236 (SD = 0.272) corresponding to a small effect sensu Cohen (1988). Note that the associated 95% CIs are generally wide, and thus the precision of effect size estimates is limited. I suspect that this shortcoming will be common in the study of behavioral variables. For definition of variables and their measurements, see Garamszegi et al. (2004) and Török et al. (2003).

[a] The sign of effects are adjusted to a direction in which higher quality males produce lower frequency minimum (e.g., a negative correlation represents a positive sign under this theoretical consideration).

[b] Means (±SDs) are for the raw variables without considering age effects, but prior to the calculation of effect sizes and their analyses, wing patch size was standardized across yearlings and adults.

direction may, however, indicate a more general role for the predictor variables affecting the response variables. The estimation of the mean effect size may allow us to decide whether only strong (i.e., significant) effects are important or weak effects also play, at least partially, a directional role (see an example in Table 1). Note that a test for consistent directional patterns requires the sign of the effects to be adjusted carefully. Directional conclusions are biologically relevant only if the variables are defined in such a way that larger values always indicate higher elaborateness or reflect superior quality. Making such directional definitions is sometimes difficult as the observer may not know a priori whether the expression of large or small values is selectively advantageous (see an example in Table 1). This problem can be treated by the theoretical reconsiderations about the direction of particular relationships and the subsequent adjustment of the signs of the effects or by the omission of the problematic variables from the mean statistics (see also Garamszegi et al. 2006 for an alternative solution).

Second, effect sizes and the corresponding CIs may stimulate meta-analytic thinking (Thompson 2002). The magnitude of an effect can be assessed from the available sample with certain precision, and thus effect size estimates are always associated with CIs. Although standardized effect sizes appear comparable regardless of sample size by definition, treating effect sizes that have different CIs in the same way may be misleading. In fact, when sample sizes are small, sampling errors can bias effect size estimations (see Chow 1988, 1998; Thompson 2002), which renders comparisons of effect sizes conservative. However, meta-analytical techniques offer quantitative methods to examine the magnitude and the generality of a predicted relationship while taking sample sizes and thus CIs into account (Hedges and Olkin 1985; Rosenthal 1991; Cooper and Hedges 1994). In such an approach, based on the standardized effect sizes and the associated CIs, an overall effect size may be calculated for the relationship in focus, and the general significance of the studied phenomenon can be tested. Accordingly, mean effect size and its CI from a meta-analysis may reflect better the true effects of a predictor variable on a set of response variable than mean effect size taken from the distribution of particular effect sizes. The results of meta-analyses can be graphically illustrated (e.g., Figure 1), and this allows the readers to visually assess the magnitude, direction, and generality of different effects, as well as the precision of their estimates. Additionally, meta-analytical approaches involve tests of heterogeneity that tell statistically whether particular effect sizes originate from a common distribution, that is, whether they play similar or different roles in shaping the general pattern (see also Sokal and Rohlf 1995, pp. 580–3, and Figure 1 for an example).

Third, when neglecting the direction of the relationships, unsigned effect sizes can be used to reflect the strength of a given relationship, for instance, according to Cohen's (1988) conventions (from $r = 0.1$ or $d = 0.2$ for small effect to $r = 0.5$ or $d = 0.8$ for large effect). The mean of the absolute values of the effect sizes may show that weak or strong effects are at work in general without considering directional roles. It may be informative to provide information about the distribution (median, skewness, and normality) of the unsigned effect sizes. For example, one may expect that among the investigated traits, only a few will play biologically important roles, and thus small effect sizes will be associated with the majority of

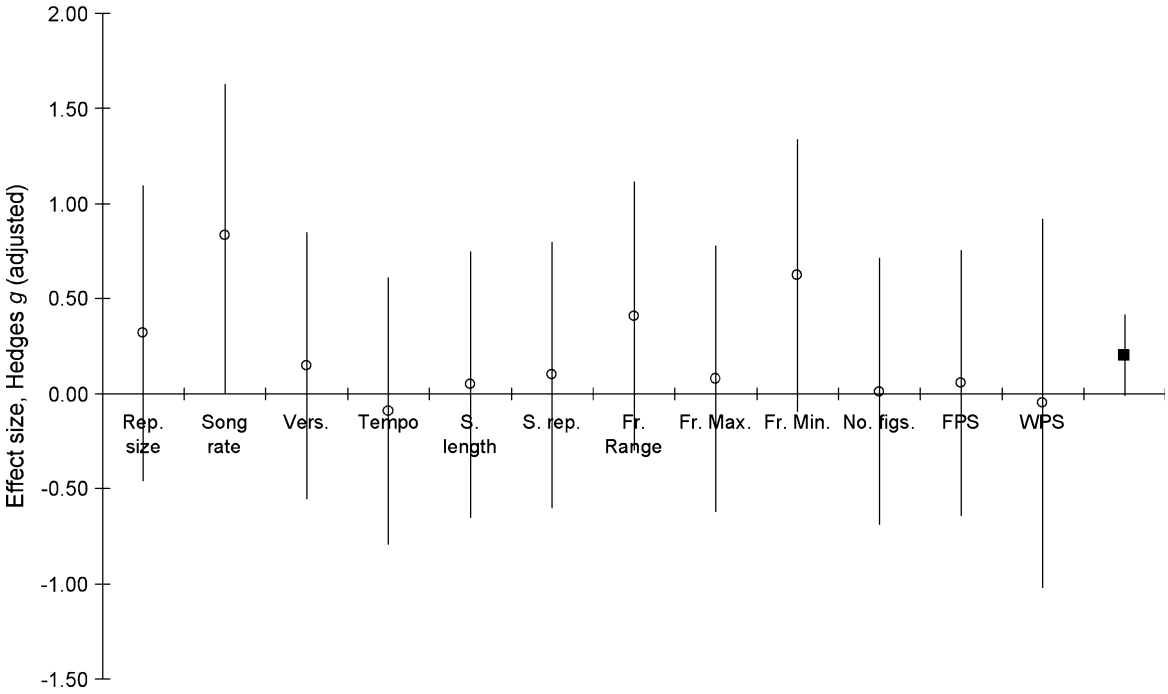

Figure 1
A meta-analysis of effect sizes (circles with error bars indicating 95% CIs) corresponding to the relationship between male sexual traits and success in nest-box retention. Note that the figure contains the information presented in Table 1. This meta-analytical approach that takes differences in CIs into account also revealed that the relationship tends to be positive (using fixed effects: Hedges' $g = 0.214$, $CI_{95\%} = 0.004$–0.423, $t = 2.008$, $P = 0.045$; using random effects: Hedges' $g = 0.202$, $CI_{95\%} = -0.009$–0.413, $t = 1.880$, $P = 0.061$, black square). This pattern seems to be homogenous across the investigated traits (test of heterogeneity: $Q = 5.943$, df = 11, $P = 0.877$).

variables resulting in a left-skewed distribution. In addition, a particular aim of many studies is to compare the strength of effects corresponding to different predictor variables, that is, to test if variable *A* or variable *B* correlates more strongly to the chosen set of variables. In this case, the absolute values of effect sizes can be used in pairwise comparisons to test whether the average strength of the relationships differs between predictor variables. In the flycatcher example (see Table 1 and Figure 1), a paired *t*-test showed no consistent difference in unsigned effect size between associations for nest-box retention and pairing success ($t_{11} = 1.045$, $P = 0.318$), implying that different components of sexual selection are linked to sexual traits with magnitudes that did not differ significantly. Furthermore, the researcher may want to compare unsigned effect sizes between different groups of traits, such as between plumage and song traits in Table 1. This would show whether plumage or song traits are important in nest-box retention reflecting male–male competition. Note that the use of unsigned effect sizes in statistical analyses while neglecting their CIs requires cautious interpretations, as explained above. Again, meta-analyses may offer partial solutions.

Fourth, if it is biologically relevant, it may be interesting to test for a relationship between the effects sizes of 2 predictor variables. If different mechanisms are responsible for the detected effects for each predictor variable, different traits with different magnitudes will be associated with the predictor variables. In this case, at the level of variables, the effect sizes should not covary between the predictor variables (see Figure 2 as an example). On the other hand, if similar mechanisms shape the observed patterns, similar relationships will be found for both predictor variables, and effect sizes may be positively associated across them. For such a test to be robust, it is important also to assess the relationship between the predictor variables themselves. It may happen that we find a correlation between effect sizes for 2 predictor variables but that this is due to a close positive association between the predictor variables (see also below).

CONFOUNDING EFFECTS: COVARIATION BETWEEN TRAITS AND LOW POWER

Effect sizes are estimated from the same sample of individuals; therefore, they are not independent observations. This nonindependence violates one of the most important assumptions of parametric tests and meta-analyses. Hence, the association between different variables at the level of individuals may confound the analyses of effect sizes at the level of variables. One potential solution may be to calculate partial correlations between the predictor variables and each of the response variables while holding the variation constant for the rest of the response variables. However, the use of such a partial correlation approach would require very complex partial correlation matrices for all variables involved with, more or less, completely filled data matrices. Unfortunately, missing values often cause difficulties in such multivariate statistics.

I suggest an alternative method to be developed that can potentially be utilized to control for the associations between variables when test statistics are based on effect sizes and variables are the unit of analysis. The relationship between different variables causes a lack of statistical independence similar to the one that arises from the use of species values as independent data points in comparative analyses (Felsenstein 1985). In comparative studies, phylogenetic approaches are applied to eliminate such confounding effects due to common ancestry to ensure statistical independence (Harvey and Pagel 1991). Being an analogous problem, similar approaches can

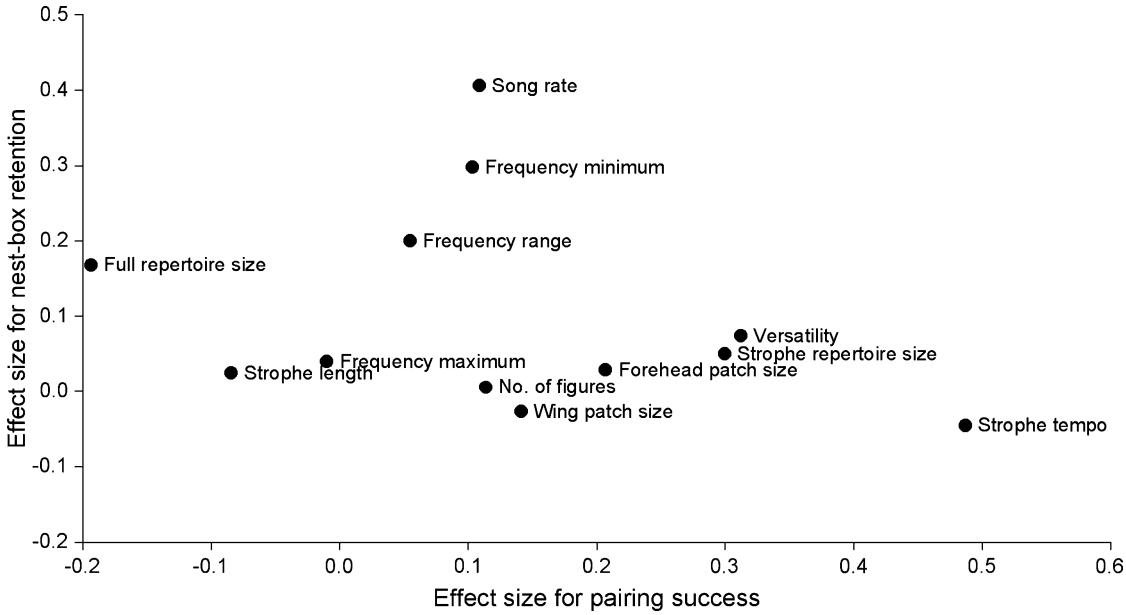

Figure 2
Due to the dual function of signals in sexual selection, in a correlative study, it is difficult to determine whether the mating success associated with a given trait is the result of female preference, or it is the consequence of male–male competition (Searcy and Andersson 1986; Berglund et al. 1996). If success in nest-box retention leads to pairing success (estimated as relative pairing date) in the collared flycatcher, the same sexual signals should be associated with them to a similar magnitude, which should generate a correlation between effect sizes for the 2 measures of mating success. However, the relationship between effect sizes of the nest-box retention/sexual signals association and between effect sizes of the pairing success/sexual signals association was not significant (signed effect sizes: $r = -0.311$, $N = 12$, $P = 0.325$; unsigned effect sizes: $r = -0.261$, $N = 12$, $P = 0.413$), even when I controlled for the covariation between sexual traits as depicted in Figure 3 by using a phylogenetic approach based on independent contrasts (signed effect sizes: $r = -0.197$, $N = 11$ contrasts, $P = 0.562$, unsigned effect sizes: $r = -0.121$, $N = 11$ contrasts, $P = 0.722$). Across individuals, there was no association between nest-box retention and pairing success ($t_{23} = 0.300$, $P = 0.766$). Hence, it is likely that different sexual signals are involved in male–male competition and mate attraction in this species. Note that the visual inspection of the figure leads to similar conclusions. The data points are the effect size estimates for 12 variables that are given as Pearson's correlation coefficients. Data from Garamszegi et al. 2004.

be used to deal with the confounding effect arising from the associations between variables. If the association between variables can be represented by a "phenetic" tree, it could subsequently be used in a phylogenetic analysis to control for the relationships between different variables. In such a tree, tips should be the variables, and different paths and branch lengths should represent their distance and relatedness (see an example in Figure 3). The hierarchical classification of the response variables based on joining- or tree-clustering methods with a single linkage can result in such structures (Podani 2000). Tree-clustering methods use dissimilarities or distances between objects to group objects of similar kind into respective categories. As the distance between variables can be reflected by their relationship, a correlation matrix of variables could be used as a distance matrix in a cluster analysis. If the distance between 2 variables is estimated as $1 - |r|$, strongly correlating variables will be closely related to each other, and the distance between them will be small. Such distances could be computed for all pairwise relationships. The numeric (unsigned) correlation coefficients should be used because we are interested in controlling for the strength of different associations neglecting the direction of the patterns. Therefore, relying on the correlation of traits, one can create a distance matrix for the variables that can be used in a cluster analysis to classify variables hierarchically. The resulting tree that holds information about the relatedness of variables can subsequently be imported into a phylogenetic program that eliminates the confounding effect of the relationships between observation points (see Harvey and Pagel 1991; Pagel 1999 for different approaches), that is, causing effect sizes to be independent of correlations between variables. For example, comparative analyses based on phylogenetically independent contrasts (CAIC) use the phylogeny of the species in the data set to partition the variance among species into independent comparisons (so-called linear contrasts), each comparison being made at a different node of the phylogeny (Purvis and Rambaut 1995). The resulting contrasts can be analyzed validly in standard statistical packages to test hypotheses about correlated evolution of traits. Similarly, based on the estimated effect sizes, independent contrasts can be calculated for each node of the phenetic tree of variables (such as in Figure 3), and these contrasts can be used to test hypotheses about the strength of relationship between different biological effects.

Note that despite the analogies, CAIC was especially developed for phylogenetic analyses and may be sensitive to specific assumptions. The applicability of the phylogenetic framework in the current context should be tested in the future, and specific methods may be developed to deal with the nonindependence of effect sizes. Until then, analyses of effect sizes should be interpreted with caution. However, generalizations by graphical approaches, such as the distribution of effect sizes, meta-analytical summaries, or the phenetic tree of variables, could already provide us with important biological information.

An additional problem may appear when test statistics are based on effect sizes. Because these approaches use variables as the unit of analysis, the sample size will be equal to the number of variables involved. Therefore, the power of the suggested tests may be limited, and conclusions based on the associated P values will be sensitive to Type II errors. In fact, below a certain limit, making analyses at the level of variables does not make much sense. When only a few variables are

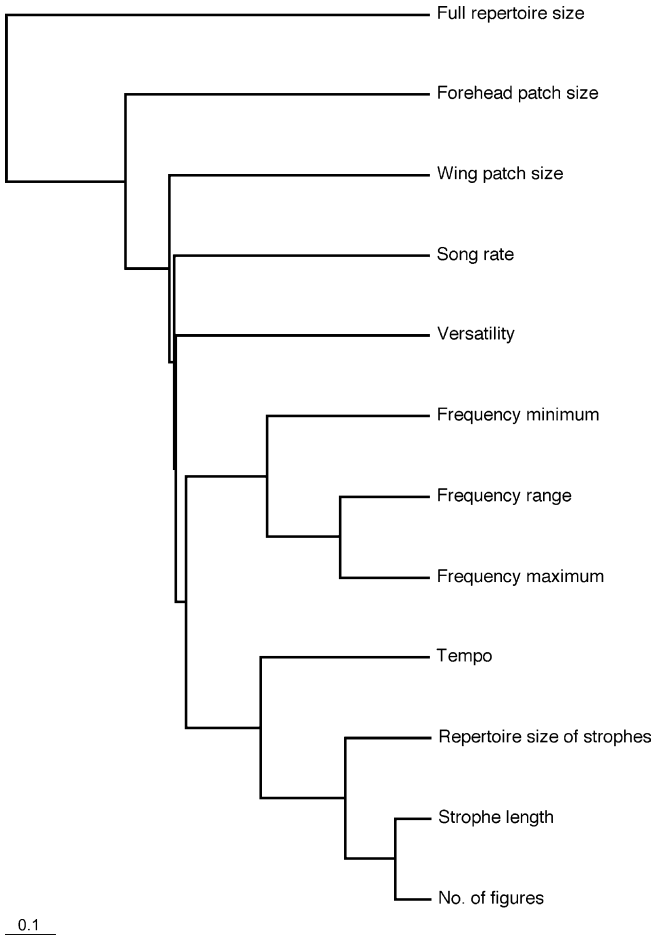

Figure 3
Hierarchical classification of 12 different sexual traits based on pairwise correlation coefficients (r) between the variables calculated for 34 male collared flycatchers. First, I calculated the pairwise correlation of traits from which I built a correlation matrix. Second, because I focused on the strength of the relationships irrespective of their direction, I removed the signs of each correlation and expressed the "phenetic" distance of traits as $1 - |r|$. Hence, small values reflect strong associations, that is, represent closely related variables. Third, this distance matrix was used in a cluster analysis to hierarchically classify sexual traits based on joining methods with a single linkage. The resulting phenetic tree of the variables is presented, which was entered in a phylogenetic program to remove the relatedness of sexual traits when they were used as the units of analysis (see Figure 2). The unit for the tree is $1 - |r|$ reflecting the distance between traits, and it is given at the bottom left.

considered, the explanation of individual effect sizes (and CIs) should be preferred. However, as the number of variables increases, the suggested analyses become more powerful, corresponding to the increased need to be able to make generalizations. In these situations, I would avoid focusing merely on significance levels and thus committing the same errors again. The framework involving graphical approaches outlined above has the potential to capture biological patterns requiring no statistical tests of significance.

CONCLUSION

In a stimulating paper, Nakagawa (2004) urged that instead of the selective presentation of results after Bonferroni correction, effects sizes (and corresponding CIs) from multiple tests should be fully presented to avoid publication bias and false interpretations in behavioral ecology. Here, I suggest that simple analyses of standardized effect sizes may further help us to understand general patterns. The recommended analyses have the potential to assess the relevance of weak, but biologically important effects, and allow generalizations. Such an approach can motivate researchers to present a more complete picture of their data instead of selectively discussing (or publishing) only a subset of significant results. In addition, I proposed a novel method to control statistically for correlations among variables that are otherwise treated as statistically independent observations. Although the conclusions of the example analysis did not change as a consequence of this exercise, I believe that the lack of statistical independence of observations may pose a serious problem in many analyses that are based on individual variables as observations and that neglect weak effects.

The analytical tool I presented can be used to address various biological questions, even within and between species, as effect sizes can be calculated and tabulated according to the problem at hand (see an example in Garamszegi et al. 2006). Here, I provided an example by using real data from the collared flycatcher. I showed that relying on Bonferroni adjustment, the traditional analysis of the available data would suggest that there is no relationship between the expression of sexual signals and a measure of male–male competition. However, analyses at the level of effect sizes demonstrated that the expression of 12 sexual traits appears to covary with nest-box retention in the same direction (Table 1 and Figure 1). Mean effect size reveals that generally, males with elaborate sexual signals appear more successful in nest-box retention than males with less elaborated signals confirming the prediction of sexual selection. However, analyses of unsigned effect sizes showed that the strength of this relationship is generally weak, which could be estimated with broad CIs in the current study. As there was no relationship between effect sizes for nest-box retention and pairing success (Figure 2), the 2 measures of mating success are independent components of sexual selection. These findings provide us with biologically relevant and general conclusions without the need for additional data or the drawback of creating publication bias by selective reporting of results.

Three anonymous referees provided stimulating criticism that significantly improved the manuscript and for which I am extremely grateful. I am highly indebted to M. D. Jennions for his constructive comments. J. Podani provided help with the hierarchic classification of variables. During this study, I received a postdoctoral fellowship from the Fonds voor Wetenschappelijk Onderzoek Flanders (Belgium).

Address correspondence to L.Z. Garamszegi. E-mail: laszlo.garamszegi@ua.ac.be.

Received 24 July 2005; revised 16 March 2006; accepted 22 March 2006.

REFERENCES

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57:289–300.

Berglund A, Bisazza A, Pilastro A. 1996. Armaments and ornaments: an evolutionary explanation of traits of dual utility. Biol J Linn Soc 58:385–99.

Cabin RJ, Mitchell RJ. 2000. To Bonferroni or not to Bonferroni: when and how are the questions. Bull Ecol Soc Am 81:246–8.

Chandler CR. 1995. Practical considerations in the use of simultaneous inference for multiple tests. Anim Behav 49:524–7.
Chow SL. 1988. Significance test or effect size. Psychol Bull 103:105–10.
Chow SL. 1998. Precis of statistical significance: rationale, validity, and utility. Behav Brain Sci 21:169–94.
Cohen J. 1988. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
Cooper H, Hedges V. 1994. The handbook of research synthesis. New York: Russell Sage Foundation.
Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat 125:1–15.
Garamszegi LZ, Hegyi G, Heylen D, Ninni P, de Lope F, Eens M, Moller AP. 2006. The design of complex sexual traits in male barn swallows: associations between signal attributes. J Evol Biol 10.1111/j.1420-9101.2006.01135.x.
Garamszegi LZ, Møller AP, Török J, Michl G, Péczely P, Richard M. 2004. Immune challenge mediates vocal communication in a passerine bird: an experiment. Behav Ecol 15:148–57.
Garcia LV. 2003. Controlling the false discovery rate in ecological research. Trends Ecol Evol 18:553–4.
Garcia LV. 2004. Escaping the Bonferroni iron claw in ecological studies. Oikos 105:657–63.
Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.
Hedges LV, Olkin I. 1985. Statistical methods for meta-analysis. London: Academic Press.
Koricheva J. 2003. Non-significant results in ecology: a burden or a blessing in disguise? Oikos 102:397–401.
Møller AP, Jennions MD. 2002. How much variance can be explained by ecologists and evolutionary biologists. Oecologia 132:492–500.
Møller AP, Thornhill R, Gangestad SW. 2005. Direct and indirect tests for publication bias: asymmetry and sexual selection. Anim Behav 70:497–506.
Moran MD. 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. Oikos 102:403–5.
Nakagawa S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15:1044–5.
Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–84.
Perneger TV. 1998. What's wrong with Bonferroni adjustments. Br Med J 316:1236–8.
Podani J. 2000. Introduction to the exploration of multivariate biological data. Leiden, The Netherlands: Backhuys Publishers.
Purvis A, Rambaut A. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Comp Appl Biosci 11:247–51.
Rice WR. 1989. Analysing tables of statistical tests. Evolution 43:223–5.
Rosenthal R. 1991. Meta-analytic procedures for social research. Thousand Oaks, CA: Sage Publications.
Searcy WA, Andersson M. 1986. Sexual selection and the evolution of song. Annu Rev Ecol Syst 17:507–33.
Sokal RR, Rohlf FJ. 1995. Biometry. 3rd ed. New York: W. H. Freeman & Co.
Thompson B. 2002. What future quantitative social science research could look like: confidence intervals for effect sizes. Educ Res 31:25–32.
Török J, Hegyi G, Garamszegi LZ. 2003. Depigmented wing patch size is a condition-dependent indicator of viability in male collared flycatchers. Behav Ecol 14:382–8.
Wilkinson L, Task Force Stat Inference. 1999. Statistical methods in psychology journals: guidelines and explanations. Am Psychol 54:594–604.
Wright SP. 1992. Adjusted P-values for simultaneous inference. Biometrics 48:1005–13.