A survey of the statistical power of research in behavioral ecology and animal behavior

Michael D. Jennions[a] **and Anders Pape Møller**[b]
[a]School of Botany and Zoology, Australian National University, Canberra, A.C.T. 0200, Australia,
Smithsonian Tropical Research Institute, Apartado 2072, Balboa, Republic of Panama, and
[b]Laboratoire d'Ecologie Evolutive Parasitaire, CNRS FRE 2365, Université Pierre et Marie Curie,
Bât. A, 7ème étage, 7 quai St. Bernard, Case 237, F-75252 Paris Cedex 5, France

We estimated the statistical power of the first and last statistical test presented in 697 papers from 10 behavioral journals. First tests had significantly greater statistical power and reported more significant results (smaller p values) than did last tests. This trend was consistent across journals, taxa, and the type of statistical test used. On average, statistical power was 13–16% to detect a small effect and 40–47% to detect a medium effect. This is far lower than the general recommendation of a power of 80%. By this criterion, only 2–3%, 13–21%, and 37–50% of the tests examined had the requisite power to detect a small, medium, or large effect, respectively. Neither p values nor statistical power varied significantly across the 10 journals or 11 taxa. However, mean p values of first and last tests were significantly correlated across journals ($r = .67$, $n = 10$, $p = .034$), with a similar trend for mean power ($r = .63$, $n = 10$, $p = .051$). There is therefore some evidence that power and p values are repeatable among journals. Mean p values or power of first and last tests were, however, uncorrelated across taxa. Finally, there was a significant correlation between power and reported p value for both first ($r = .13$, $n = 684$, $p = .001$) and last tests ($r = .16$, $n = 654$, $p < .0001$). If true effect sizes are unrelated to study sample sizes, the average true effect size must be nonzero for this pattern to emerge. This suggests that failure to observe significant relationships is partly owing to small sample sizes, as power increases with sample size.
Key words: effect size, meta-analysis, publication bias, sample sizes, statistical power. *[Behav Ecol 14:438–445 (2003)]*

The biological literature is dominated by reports of statistically significant patterns of association (Csada et al., 1996). This may partly reflect a publication bias toward significant findings (Kotiaho and Tomkins, 2002; Palmer, 1999, 2000). Recent reviews show that studies with both small sample sizes and nonsignificant results are underrepresented in the literature (Jennions and Møller, 2002a,b). This could bias our assessment of the average strength of biological relationships. Biologists need to ensure that studies are equally publishable whether their results are significant or not. This, however, raises a problem. Should the criteria for publication be an a priori minimal level of confidence in the conclusion of a study in case the observed outcome turns out to be a nonsignificant result? If the answer is yes, then, to determine publishability, we must calculate our confidence in a conclusion that there is no significant effect. If the answer is no, we must still do this because nonsignificant results are then published, and readers need to assess how much confidence to place in a negative conclusion. Leaving aside whether a dichotomy into significant and nonsignificant results is appropriate for biologists (Stoehr, 1999), this question is best answered by statistical power analysis (but see Hoenig and Heisey, 2001). Power is the probability of obtaining a significant result when the null hypothesis is false. Power increases as sample size, α-level of significance, and effect size (magnitude of the difference between the alternative and null hypothesis) increase, and decreases with greater variance in the study population. If the effect size is a standardized measure (e.g., the mean difference between two groups expressed in standard deviations, d; or the correlation coefficient, r), it is dimensionless, and there is no need to specify population variance to calculate statistical power (Thomas and Krebs, 1997). The use of standardized measures can, however, yield differences in observed effect size solely owing to predictable differences in the likelihood of measurement error (e.g., between laboratory and field studies; Hurlbert, 1994). In general, however, biologists report post-hoc statistical power to detect standardized measures of effect sizes of specific magnitude, conventionally referred to as small, medium, or large effects (Cohen, 1988).

Despite being urged to incorporate power analysis into research design and presentation (see Greenwood, 1993; Peres-Neto and Olden, 2001; Stoehr, 1999; Thomas and Juanes, 1996; Thompson and Neill, 1993; Toft and Shea, 1983), most behavioral ecologists still report nonsignificant results without indicating a test's statistical power (Stoehr 1999; this study). Since the first survey by Cohen (1962), those in the medical and social sciences have conducted surveys to estimate average power in specific fields or journals (see Chung et al., 1998; Kloster and Layne, 1997; Maddock and Rossi, 2001; Moher et al., 1994; additional examples in other disciplines are given by Cohen, 1988: xi). In biology, the effect of low power has been examined in a few specific areas. For example, Noor and Smith (2000) showed that low power might affect conclusions of studies on sexual isolation in *Drosophila.* Palmer (2000) recently pointed out that the ability to detect small to medium deviations from a one-to-one sex ratio in vertebrates with even moderate power requires sample sizes far larger than those in most published studies. To our knowledge, there has been no systematic attempt to conduct a power analysis survey of a broad area of research in biology.

Why quantify *average* statistical power? After all, interpreting the results of a specific statistical test depend solely on its power. We propose four reasons. First, surveys invariably show that the general power to detect relationships is far lower than most researchers think (see Dickinson et al., 2000; Kazantzis,

Address correspondence to M.D. Jennions at the School of Botany and Zoology, Australian National University, Canberra, A.C.T. 0200, Australia. E-mail: michael.jennions@anu.edu.au.

Received 7 November 2001; revised 25 June 2002; accepted 2 September 2002.

2000). Ignorance of the relationship between sample size and power could explain why researchers often conduct studies with small sample sizes, when even modest increases could have greatly improved their statistical power (Thomas and Juanes, 1996). When confronted with the reality of low power, researchers may be encouraged to explicitly consider sample size and improve experimental design before conducting studies. Second, power surveys can determine whether researchers are becoming more aware. If they are, statistical power should increase through time. For example, Rossi (1990) and Sedlmeier and Gigerenzer (1989) replicated Cohen's original 1962 survey and found no increase in power between 1960 and 1982–1984. In contrast, Moher et al. (1994) reported an increase in power in randomized control trials in medicine over a 25-year period. Stoehr (1999) disagreed with a reviewer who felt he had created a straw man, by stating that most behavioral researchers do not take power analysis seriously. This difference in opinion is easily resolved by comparing the mean power of tests conducted now and in the future (or past). Third, if the literature is replete with tests with low power, this should influence how researchers interpret the literature. (Even if the main focus of a study is a clearly significant result, interpretation often relies on the absence of confounding variables. Their absence is usually based on subsidiary tests reporting nonsignificant results.) Many papers in behavioral ecology reach strongly worded conclusions after refuting an alternate hypothesis, even though the power to reject the null hypothesis was extremely low. This leads to fallacious vote counting if scientists simply tally the proportion of studies that detect a relationship without considering the influence of sample size (Cooper and Hedges, 1994). Four, behavioral ecologists may fail to report power for fear that this will reduce the likelihood of publication when reviewers see low values. The review process, however, involves assessment of manuscripts *relative* to hypothetical alternatives (e.g., is this paper in the top 25%?). These alternatives are unavailable, so a reviewer's assessment is usually based on the quality of previous studies. Knowing the statistical power of recently published studies will provide an empirical benchmark that allows better informed decisions. A statistical power of 30% to detect a small effect is actually impressively high when compared with the average.

Here we present a power survey of papers from 10 journals. Eight of these are devoted to ethological and behavioral ecological studies. The other two often contain studies directly reporting on animal behavior or the immediate consequences thereof (e.g., effect of habitat choice on spatio-temporal abundance).

METHODS

Analysis of statistical power requires knowledge of the effect size we wish to detect. For standardized measure of the magnitude of a relationship, Cohen (1988) has defined small, medium, and large effect sizes for several tests. For example, a small effect has a mean correlation coefficient, *r,* of .10 (i.e., explains 1% of the variance because $r^2 = 1\%$), a medium effect has $r = .30$, and a large effect has $r = .50$. Biologists usually perform analyses estimating the power to detect an effect of medium strength. Here we present data on statistical power if the effect size is small, medium, or large at the $p = .05$ level (two-tailed) as defined in chapters 2–8 of Cohen (1988). However, a recent survey of 44 biological meta-analyses examining 242 null hypotheses shows an average effect for ecological or evolutionary studies of $r = .18$–$.19$ (Møller and Jennions, 2002). Thus, the observed average effect size examined by biologists (at least for relationships subjected to meta-analysis) is below the medium effect of Cohen (1988). We do not know the true effect sizes for the tests we examined. As such, we cannot conclude there is a publication bias simply because most studies report significant results (Bauchau, 1997). The expected proportion of studies reporting significant results depends on the effect size these studies were trying to detect (and the sample sizes).

We estimated power for 1362 statistical tests from 697 original papers from 10 journals: *American Naturalist* (33), *Animal Behaviour* (187), *Behaviour* (69), *Behavioral Ecology* (68), *Behavioral Ecology and Sociobiology* (102), *Behavioural Processes* (39), *Ethology* (73), *Journal of Insect Behaviour* (54), *Journal of Animal Ecology* (50), and *Ethology, Ecology and Evolution* (22). The number of usable papers per journal is indicated in parenthesis. In each case, we examined all issues of the journal with a 2000 publication date (only the Nov/Dec issue of *Behaviour* was unavailable).

For each paper, we looked for the first and last statistical test presented in the text of the Results section. We defined a statistical test as having been presented if the author(s) reported a probability value (henceforth, *p* value) either exactly or using the phrase "$p < .0X$" or "$p > .Y$," or if the shorthand "N.S" or "n.s." was used to denote a nonsignificant *p* value, and it was clear which statistical test had been used. We did not consider a test to have been presented if the authors simply made a statement such as "there was no difference between *X* and *Y*" or "there was a significant correlation between *X* and *Y*." If there were fewer than two usable statistical tests in the main text, we then looked at tables and figure legends, reading from top left to bottom right. For 665 of the 697 papers, we obtained data for two tests. Use of the first and last test provided an objective way to collect data. Focusing on the so-called main test of a study may be misleading. Most papers emphasize statistically significant findings, even though these may not have been the original focus of the study (Csada et al., 1996; but see Bauchau, 1997).

For each test, we recorded the *p* value. If *p* was given as $p < X$, we set $p =$ X. If given as $p > Y$, we only set *p* as Y if Y $> .05$. (In 24 of 1362 tests [1.8%], the only information was that $p > .05$.) For summary analyses, we converted *p* values into their associated standard normal deviates (*z* scores, $z = 1.96$ when $p = .05$). We also recorded the taxa/type of study using 11 categories: crustaceans, insects, spiders, other invertebrates (excluding insects, spiders, and crustaceans), fish, amphibians, reptiles, birds, mammals, plants, and species level (e.g., phylogenies). We analyzed mean power for different taxa because there is a general assumption that studies of some taxa (e.g., mammals) have smaller sample sizes than others (e.g., insects) and that these will therefore have reduced statistical power.

We calculated the power of the most commonly encountered statistical tests, specifically binomial tests; sign tests; χ^2 goodness-of-fit or R × C contingency table analyses; *G* tests (log-likelihood ratio tests for contingency tables); comparison of two proportions; comparison of two correlation coefficients; Fisher's Exact test for a 2 × 2 table; Friedman's nonparametric test; Kruskal-Wallis nonparametric one-way ANOVA; Mann-Whitney *U* test (two independent samples); paired *t* test; tests for significance of correlation coefficients; one-sample *t* test; two-sample *t* test; Wilcoxon's matched-pairs test; one-way ANOVA; and test of main effects for fixed factors in ANOVA with simple factorial designs (two-way or three-way ANOVAs), which includes tests for difference in elevation in ANCOVAs (Cohen, 1988: 379). The only tests excluded with any regularity were tests for main effects in ANOVAs with complex designs (specifically repeated measures and nested factors), tests for interaction terms in all ANOVAs, logistic regressions, and tests based on maximum likelihood or restricted maximum likelihood approaches. These tests were mainly excluded because of ease of power analysis. They were not excluded on a priori evidence that they had lower, or higher, statistical power than that of the included tests.

Statistical power was calculated using tables in Cohen (1988). In addition, we used G*power to calculate power for sample sizes smaller than those presented in the tables (Erdfelder et al., 1996). G*power and Cohen's tables show close agreement, with the exception of F tests in which the approximation method of Cohen can lead to differing results with complex experimental designs (Bradley et al., 1996). However, we only used Cohen (1988) to calculate power for simple one-way, two-way, and three-way ANOVAs. F tests of the significance of a regression were analyzed as tests of the significance of the regression coefficient. This is equivalent to testing the significance of the correlation coefficient (Cohen, 1988: 76–77). To allow us to collect and analyze data on so many tests, we made a few simplifying assumptions that slightly *inflated* our estimates of power. They were as follows.

1. For four nonparametric tests, we calculated the power if the available data had been analyzed using the equivalent parametric test. The relative power (efficiency) of nonparametric tests is weaker than their parametric counterparts because they make fewer assumptions (Siegel and Castellan, 1988: 21). However, with moderate to large sample sizes, the power of nonparametric tests becomes similar to that of the equivalent parametric tests. Specifically, for Wilcoxon tests we calculated power for a paired t test; for Mann-Whitney U test for a two-sample t test; and for Kruskal-Wallis and Friedman's test for parametric ANOVAs. For Mann-Whitney U tests, Wilcoxon tests, and Kruskal-Wallis tests, the statistical power reaches about 95.5% of that of the equivalent parametric t tests for moderate sample sizes. For Friedman's test, power is 64% of that of the equivalent F test when there are two groups, increasing to 87% for five groups (Siegel and Castellan, 1988). The estimated power we report for these nonparametric tests is therefore slightly larger than the actual power of the tests used by the original authors.
2. For two nonparametric tests, we calculated power for equivalent nonparametric tests. For G tests we calculated power for χ^2 tests. These two methods of analysis usually yield the same conclusions, and there is no clear agreement as to which is preferable (Zar, 1999: 475), so they tend to be used interchangeably by researchers. For Fisher's Exact test, we calculated power for a comparison of two proportions. This was the hypothesis tested by the original authors as one of the margins has fixed totals (Cohen, 1988: Table 6.3.5).
3. In some two-sample t tests or Mann-Whitney U tests, only the total sample size was given. We assumed that group sample sizes were equal, which maximizes power. Again, this will slightly inflate our estimates.

We compared statistical power and z scores among journals, taxa, and statistical test types by using Kruskal-Wallis one-way ANOVA. (Power for the equivalent parametric ANOVA to detect small, medium and large effects is given in parenthesis. In all cases, parametric tests yielded the same conclusions.) We compared first and last tests by using Wilcoxon's tests. When comparing power among groups, we used "power to detect a medium effect" as the dependent variable. This is likely to maximize the difference between groups because power is a percentage that shows asymptotic values at small and larger sample/effect sizes. Following the method of Stoehr (1999), we also present the results of our statistical tests as observed effect size, by converting the test statistic or p value to r following the method of Cooper and Hedges (1994: 236–240). To avoid confusion, we denote these as E_r. Rather than presenting the mean effect size, we present the 95% confidence intervals. These provide the clearest indication of the certainty with which we can conclude that an effect differed from zero (for an excellent review, see Hoenig and Heisey, 2001; see also our Discussion).

Table 1

Three power estimates and z scores for first and last statistical tests

	First test	Last test
z score	2.30 ± 0.04 (684)	1.89 ± 0.04 (654)
Power (small) (%)	16.2 ± 0.68 (697)	12.8 ± 0.56 (665)
Power (medium) (%)	47.2 ± 1.13 (697)	39.4 ± 1.05 (665)
Power (large) (%)	72.3 ± 1.00 (697)	65.3 ± 1.04 (665)

Mean ± SE. Sample sizes are in parentheses.

RESULTS

A few papers presented statistical power for the "main" biological hypothesis under test. However, power was not reported in any of the 533 tests with nonsignificant results. We had to calculate power by using the reported sample size. In many instances, even this was difficult. A close inspection of the paper was sometimes required to track down the information (e.g., *Journal of Animal Ecology* in which sample sizes were presented in the Methods section but not in the Results section). In some cases, we were unable to work out the sample size, either because of an ambiguity in the paper or because it was never provided.

First and last tests

There were clear differences between first and last statistical tests. First tests had significantly greater statistical power (Wilcoxon's test: $n = 550$ pairs, $z = 6.72$, $p < .0001$; $E_r \approx 0.208$–0.362) and smaller p values (Wilcoxon's test: $n = 584$ pairs, $z = 7.43$, $p < .0001$; $E_r \approx 0.135$–0.290); 68.4% of the 697 cases for first tests and 52.8% of the 665 cases for last tests were significant at the 0.05 level (Table 1). Across papers, however, the p values and power of first and last tests were both positively correlated (p values: $r = .143$, $n = 644$, $p < .0001$, $E_r = 0.066$–0.218; power: $r = .427$, $n = 665$, $p < .0001$, $E_r = 0.363$–0.487). Based on a comparison of average power and p values of first and last tests, this trend was consistent across journals ($z = 2.80$, $p = .005$, $E_r = 0.586$–0.973; $z = 2.701$, $p = .0069$, $E_r = 0.485$–0.965, both $n = 10$), taxa ($z = 2.93$, $p = .0033$, $E_r = 0.611$–0.970; $Z = 2.76$, $p = .0058$, $E_r = 0.753$–0.983, both $n = 11$) and statistical test types ($z = 2.42$, $p = .016$, $E_r = 0.172$–0.875; $z = 3.17$, $p = .0015$, $E_r = 0.579$–0.951, both $n = 14$) (all Wilcoxon tests). The more important conclusion, however, is that statistical power is generally low. For first tests, mean power is 13–16% to detect a small size effect and 40–47% to detect a medium effect. This is far lower than the generally recommended 80% (Cohen, 1988: 56) or 95% (Peterman, 1990). Using the 80% criterion for first statistical tests, only 2.9%, 21.2%, and 49.8% of 697 cases had the requisite power to detect a small, medium, or large effect, respectively. Likewise, for second tests, only 1.8%, 13.2%, and 36.5% of the 665 cases had sufficient power. If we only consider those tests that reported nonsignificant relationships, the equivalent figures are 1.4%, 17.8%, and 47.0% for the 219 first tests and 1.3%, 10.8%, and 32.5% for the 314 second tests.

Table 2

***z* scores and power to detect a medium effect for 10 biological journals**

Journal	*z* score		Power (medium)	
	(First test)	(Last test)	(First test)	(Last test)
American Naturalist	2.67 ± 0.19 (33)	2.33 ± 1.88 (32)	45.0 ± 5.2 (33)	42.0 ± 4.8 (32)
Animal Behaviour	2.33 ± 0.08 (184)	1.88 ± 0.08 (176)	46.8 ± 2.2 (187)	37.6 ± 2.0 (179)
Behavioral Ecology	2.13 ± 0.13 (68)	1.66 ± 0.13 (67)	51.2 ± 3.6 (68)	41.0 ± 3.3 (67)
Behavioural Ecology and Sociobiology	2.25 ± 0.11 (99)	1.75 ± 0.11 (98)	52.9 ± 2.9 (102)	42.2 ± 2.7 (99)
Behavioural Processes	2.38 ± 0.17 (39)	1.98 ± 0.17 (38)	40.0 ± 4.8 (39)	33.5 ± 4.4 (38)
Behaviour	2.45 ± 0.13 (67)	2.05 ± 0.14 (63)	44.6 ± 3.6 (69)	40.2 ± 3.4 (63)
Ethology, Ecology, and Evolution	2.15 ± 0.23 (22)	2.13 ± 0.24 (21)	58.0 ± 6.3 (22)	44.0 ± 5.9 (21)
Ethology	2.24 ± 0.13 (73)	2.03 ± 0.13 (66)	43.0 ± 3.5 (73)	35.5 ± 3.3 (69)
Journal of Animal Ecology	2.28 ± 0.15 (49)	1.73 ± 0.16 (45)	49.3 ± 4.2 (50)	40.3 ± 4.0 (45)
Journal of Insect Behaviour	2.23 ± 0.15 (50)	1.85 ± 0.16 (48)	42.2 ± 4.0 (54)	43.1 ± 3.8 (52)

Mean ± SE. Sample sizes are in parentheses.

Journal, taxa, and test type

Neither p values (first test: $\chi^2 = 10.48$, $p = .313$; last test: $\chi^2 = 14.68$, $p = .10$) nor statistical power (first test: $\chi^2 = 15.61$, $p = .08$; last test: $\chi^2 = 6.78$, $p = .66$) varied significantly among the 10 journals (all Kruskal-Wallis tests, df = 9; Table 2; power for one-way ANOVA: all >37%, >99.5%, >99.5%). This suggests that neither impact factor, policy of the journal or any other assessment of journal "quality" is related to the statistical significance of the results presented or the sample sizes on which conclusions are based.

Neither p values (first test: $\chi^2 = 9.07$, $p = .53$; last test: $\chi^2 = 7.59$, $p = .67$) nor statistical power (first test: $\chi^2 = 9.58$, $p = .48$; last test: $\chi^2 = 11.84$, $p = .30$) varied significantly among the 11 taxa (all Kruskal-Wallis tests, df = 10; Table 3; power for one-way ANOVA: all >35%, >99.5%, >99.5%). We then reanalyzed the data only looking at the three taxa with large sample sizes (birds, mammals, and insects). Again, there was no difference among taxa in p values (first test: $\chi^2 = 2.30$, $p = .32$; last test: $\chi^2 = 2.51$, $p = .29$). For statistical power, there was no difference for the first test ($\chi^2 = 2.29$, $p = .32$), but there was for the last test ($\chi^2 = 8.28$, $p = .016$; all Kruskal-Wallis tests, df = 2; power for one-way ANOVA: all >46%, >99%, >99%). Bird studies had less power than insect studies ($p < .05$ post-hoc pair-wise comparison).

Among the 14 statistical test types, p values did not differ significantly (first test: $\chi^2 = 15.69$, $p = .27$; last test: $\chi^2 = 12.90$, $p = .46$); but statistical power did (first test: $\chi^2 = 95.45$, $p < 0.0001$; last test: $\chi^2 = 70.86$, $p < .0001$; all Kruskal-Wallis tests, df = 13; Table 4; power for one-way ANOVA: >36%, >99.5%, >99.5%).

Mean p values of first and last tests were significantly correlated across journals ($r = .670$, $n = 10$, $p = .034$, $E_r = 0.070–0.914$), as was mean power ($r = .630$, $n = 10$, $p = .051$, $E_r = 0.001–0.902$). There is therefore some evidence that power and p values are repeatable among journals. Mean p values or power of first and last tests were not correlated across taxa ($r = .232$, $p = .493$, $n = 11$, $E_r = -0.394–0.711$; $r = .18$, $p = .596$, $n = 11$, $E_r = -0.439–0.683$; power: 6%, 16%, 40%); nor were first and last test p values correlated across statistical tests type ($r = -.186$, $p = .525$, $n = 14$, $E_r = -0.652–0.382$; power: 6%, 18%, 47%), although mean power was ($r = .974$, $p < .0001$, $n = 14$, $E_r = 0.917–0.992$).

Power and p values

There was a significant positive correlation between power and the reported p value (expressed as z score) for both first ($r = .125$, $p = .001$, $n = 684$, $E_r = 0.051$ to 0.198) and last tests ($r = .162$, $p < .0001$, $n = 654$, $E_r = 0.086$ to 0.236). Thus, the greater the statistical power, the more often the test reported a significant effect. To determine whether this relationship was owing to combining different types of tests,

Table 3

***z* scores and power to detect a medium effect for 11 taxonomic categories**

Taxa	*z* score		Power (medium)	
	(First test)	(Last test)	(First test)	(Last test)
Crustaceans	2.00 ± 0.25 (18)	1.96 ± 0.26 (17)	52.1 ± 7.0 (18)	32.8 ± 6.3 (18)
Spiders	2.45 ± 0.25 (19)	2.07 ± 0.26 (18)	51.5 ± 6.8 (19)	41.0 ± 6.2 (19)
Insects	2.35 ± 0.09 (152)	2.00 ± 0.09 (143)	50.2 ± 2.4 (158)	45.5 ± 2.2 (148)
Other invertebrates	2.53 ± 0.23 (21)	2.21 ± 0.24 (21)	39.9 ± 6.3 (22)	40.5 ± 5.9 (21)
Amphibians	2.20 ± 0.19 (32)	1.86 ± 0.20 (31)	40.3 ± 5.3 (32)	37.2 ± 4.8 (31)
Reptiles	2.37 ± 0.23 (21)	2.11 ± 0.24 (21)	53.2 ± 6.5 (21)	40.2 ± 5.9 (21)
Fish	2.44 ± 0.13 (64)	1.81 ± 0.14 (62)	42.5 ± 3.7 (64)	41.7 ± 3.4 (62)
Birds	2.28 ± 0.08 (185)	1.79 ± 0.08 (180)	46.1 ± 2.2 (188)	35.1 ± 2.0 (182)
Mammals	2.18 ± 0.09 (138)	1.88 ± 0.10 (130)	46.0 ± 2.5 (141)	37.7 ± 2.4 (131)
Plants	2.57 ± 0.40 (7)	1.88 ± 0.41 (7)	58.9 ± 11.2 (7)	39.8 ± 10.2 (7)
Species level	2.40 ± 0.22 (23)	1.72 ± 0.24 (21)	55.0 ± 6.2 (23)	44.6 ± 5.9 (21)

Mean ± SE. Sample sizes are in parentheses.

Table 4

***z* scores and power to detect a medium effect for 14 different statistical tests**

	z score		Power (medium)	
Test type	(First test)	(Last test)	(First test)	(Last test)
Binomial	2.37 ± 0.21 (26)	1.88 ± 0.26 (18)	35.1 ± 5.5 (26)	35.0 ± 6.2 (18)
χ^2	2.20 ± 0.11 (92)	1.82 ± 0.14 (61)	70.5 ± 2.9 (93)	58.7 ± 3.3 (62)
F test (one-way)	2.27 ± 0.12 (74)	1.94 ± 0.14 (60)	44.3 ± 3.3 (74)	37.2 ± 3.4 (60)
F test (other)	2.53 ± 0.15 (52)	1.74 ± 0.16 (48)	42.7 ± 3.8 (54)	37.5 ± 3.7 (49)
Fisher's Exact	2.56 ± 0.21 (27)	2.00 ± 0.22 (25)	40.6 ± 5.3 (28)	37.6 ± 5.1 (26)
Friedman's test	2.08 ± 0.36 (9)	2.86 ± 0.54 (4)	37.5 ± 8.9 (10)	32.3 ± 13.1 (4)
Kruskal-Wallis	2.26 ± 0.24 (20)	1.76 ± 0.25 (19)	36.5 ± 6.0 (22)	35.5 ± 6.0 (19)
Mann-Whitney	2.24 ± 0.14 (58)	2.09 ± 0.13 (69)	39.9 ± 3.9 (61)	31.9 ± 3.1 (73)
Paired *t* test	2.19 ± 0.15 (52)	1.92 ± 0.16 (44)	47.3 ± 3.9 (52)	44.1 ± 3.9 (45)
Correlation coefficient	2.47 ± 0.09 (138)	1.89 ± 0.08 (173)	42.8 ± 2.9 (138)	36.4 ± 2.0 (175)
Sign test	2.15 ± 0.34 (10)	2.29 ± 0.41 (7)	26.1 ± 8.9 (10)	26.8 ± 9.9 (7)
t test (one-sample)	2.49 ± 0.31 (12)	2.33 ± 0.29 (14)	74.9 ± 7.8 (13)	60.7 ± 7.0 (14)
t test (two-sample)	1.97 ± 0.14 (57)	1.74 ± 0.14 (58)	44.7 ± 3.7 (58)	35.8 ± 3.4 (58)
Wilcoxon's paired test	2.29 ± 0.14 (57)	1.72 ± 0.15 (52)	50.4 ± 3.7 (58)	41.9 ± 3.6 (53)

Mean ± SE. Sample sizes are in parentheses.

we reexamined this relationship for specific statistical tests. Only tests in which $n \geq 84$ were used because the power to detect a medium effect at $\alpha = 0.05$ is then greater than 80%. For correlation coefficient tests, $r_s = .076$ ($p = .376$, $n = 138$, $E_r = -0.093$–0.239) and $r_s = 0.267$ ($p < .001$, $n = 173$, $E_r = 0.123$–0.400); for χ^2 tests, $r_s = 0.299$ ($p = 0.004$, $n = 92$, $E_r = 0.102$–0.476); and for *F* tests (all types combined), $r_s = 0.411$ ($p < .001$, $n = 126$, $E_r = 0.254$–0.547) and $r_s = 0.011$ ($p = .91$, $n = 108$, $E_r = -0.178$–0.200). A significant decrease in *p* value as statistical power increased was therefore also apparent for specific statistical tests in three of five cases.

DISCUSSION

The statistical power of behavioral studies to detect relationships is low. For example, the power to detect a medium effect is about 39–47%. Only 10–20% of tests exceeded the recommended minimum criterion of 80% power (Cohen, 1988). This was true whether we considered all tests or only those reporting nonsignificant results. Authors still fail to report power for nonsignificant results (see Thomas and Juanes 1996: 859), even in journals that require them to do so. For example, since 1997, *Animal Behaviour*'s Instructions for Authors has stated, "where a significance test based on a small sample size yields a non-significant outcome, the power of the test should normally be quoted." Examination of two recent copy of *Animal Behaviour* showed, however, that of 215 statistical tests in 17 papers in which $p > 0.05$, none reported statistical power (March 2001), whereas only one of 22 empirical papers presented estimates of power in the January 2001 issue. By comparison, the power to detect medium effects in other fields (mainly medical) was 37% (Sedlmeier and Gigerenzer, 1989), 48% (Cohen, 1962), 57% (Rossi, 1990), <58% (Kazantzis, 2000), 62% (Chase et al., 1978), 71% (Polit and Sherman, 1990), and 77% (Maddock and Rossi, 2001), and the proportion of studies with a power greater than 80% to detect a medium effect was 0.36 (Moher et al., 1994), <0.50 (Chung et al., 1998), and >0.80 (Mengel and Davis, 1993). Mean statistical power in behavioral biology is therefore lower than that in medicine. These differences may be owing to the type of statistical test used, which is partly determined by experimental design (e.g., paired *t* tests are more powerful than two-sample *t* tests). They are, however, far more likely to reflect differences in sample sizes for studies conducted in the various research fields.

Stoehr (1999) noted that criticizing biologists' failure to consider power and interpret studies in terms of effect size is sometimes viewed as attacking a straw man. Surely biologists are aware of these issues? Our survey suggests otherwise. If they were, they would more often report statistical power. One solution is for editors and reviewers to ensure that all statistical tests are uniformly presented with full information on sample size (for each group if this influences power estimates as in, say, a two-sample *t* test), degrees of freedom, exact *p* values (or as precise as possible, e.g., $0.5 > p > 0.20$, but not just $p = \text{ns}$) and the statistical power to detect small and medium effects as conventionally defined (large effects are probably too rare to warrant presentation). If methods to determine power are not well established (this will be fairly rare as most behavioral papers use a limited set of well-studied statistical tests), the authors should explicitly state this in their methods. Thomas and Juanes (1996) and Thomas and Krebs (1997) review and list several free or purchasable software programs that can be used to calculate a priori statistical power.

Stoehr (1999) recommended that authors report effect sizes (hence, our own reporting of the 95% confidence intervals for E_r here). Effect sizes are easily calculated and do not require expensive software, textbooks, or heavy investment in learning complex skills. To ensure ready access to the relevant information, journals could publish in print and on their Web sites the formulae to convert common statistics such as *t*, *F* and χ^2 or *p* values to Pearson's *r*. Effect size can then be calculated by using a handheld calculator, spreadsheet, or user-friendly effect-size calculators (e.g., Metawin 2.0; Rosenberg et al., 2000). There are many effect sizes available, so it would be convenient if biologists agreed on which one to use (when possible). We suggest Pearson's *r* as the most useful because r^2 is the proportion of variance explained. It can be calculated whenever there is a directional trend (e.g., *t* tests, correlations or χ^2, or *F* tests where $\text{df} = 1$). For omnibus tests of variation among groups rather than tests for linear trends (e.g., *F* or χ^2 tests when the numerator $\text{df} > 1$), if possible R^2 should be presented. Stoehr (1999) has already listed some advantages of stating effect sizes. These are mainly related to ensuring that readers do not use *p* values when comparing the strength of relationships (which assumes equivalent sample sizes). We can add three other advantages. First, it would greatly facilitate the efforts of meta-analysts and reduce error rates when interpreting or transcribing statistical tests during a literature review (Cooper and Hedges, 1994). Second, stating effect sizes allows researchers planning to replicate or conduct similar studies to calculate more easily the sample size needed to achieve the desired statistical power. Knowing the probable effect size is a prerequisite to a well-designed experiment or data collection protocol. Three, reviewers often remind authors to report summary statistics and not just test statistics. In general, however, effect sizes may be easier to interpret. For example, stating the mean and SD for body size in two groups with different sample sizes requires the reader to somehow visualize the overlap between the groups. In contrast, stating the effect size r immediately tells the reader that group identity explains r^2 of the observed variation in body size.

If effect sizes *and* their confidence intervals are presented, power analyses gives no additional insights (Hoenig and Heisey, 2001). We have illustrated this approach here by converting our statistical results into the effect size r (written E_r) and giving the 95% confidence intervals. Confidence intervals cover a set of nonrefuted values. If these values are tightly clustered around the null value (usually zero), then we are confidence that the true value is near the null hypothesis. Conversely, a wide confidence interval, even if it includes the null value, is treated more cautiously because we know there is a good chance that the true effect size might lie far from the null value. A second advantage of the confidence interval approach is that it prevents researchers from erroneously concluding that when studies fail to reject the null hypothesis, the greater the statistical power, the greater the likelihood that the null hypothesis is true. That this is flawed reasoning is easily visualized by noting that the observed p value—hence, position of the estimated mean effect size relative to the null value (for a fixed sample size as the p value decreases effect size increases)—is also critical in determining the likelihood that the null hypothesis is correct. For example, an estimate of $r = -.02$–$.04$ is more likely to lead to the conclusion that the true effect is close to the null value of zero than a study with lower statistical power (larger confidence intervals) where $r = -.40$–$.40$.

Publication criteria

Should statistical power influence publication decisions? One view is that studies should not be published if they have low statistical power, because if they produce negative findings, there is little confidence in the oft-stated conclusion that the null hypothesis is correct. This would be fine if reviewers were equally likely to reject papers with low power that report significant results. There is, however, good evidence for a publication bias in biology toward significant results, even when sample sizes are small (Csada et al, 1996; Møller and Jennions, 2001, 2002; Palmer, 2000). This culture will be hard to eliminate. We believe it is more important that the literature as a whole is unbiased, so we take the pragmatic view that statistical power should not be a criterion for publication. In general, we think that synthesis of results from many studies is more valuable than a conclusion based on extrapolation from a few big studies. (For an excellent defense of the search for generalities in behavioral ecology rather than a focus on the peculiarities of specific systems, see Reeve, 2001.) Practically speaking, we believe that a requirement to present statistical power or confidence intervals for effect size would (1) ensure readers were fully aware of the weakness of a conclusion that there is no effect, (2) encourage researchers to increase sample sizes, and (3) be achievable without loss of print space if journals create on-line sections in which studies with negative results and low power are published. The drive to be cited and for peer acknowledgment of one's work would probably also encourage increased sample sizes as authors strove to publish in the more prestigious print section.

Specific findings

Between 53% and 68% of tests were significant at the 0.05 level. In contrast, the mean power to detect a medium strength effect was only 40–47%. Why the discrepancy? First, the true effects for the questions asked in first and last tests may be greater than $r = .3$. The estimate of a mean effect size in biology of $r = .20$ by Møller and Jennions (2002) was from meta-analyses that may deal with a different set of biological relationships. Second, mean estimates of effect size from biological meta-analyses take into consideration the direction of the effect. Thus, the mean effect is smaller than the magnitude of the absolute effect size. Third, there may be a publication bias toward significant results (Palmer, 2000). Four, authors may organize papers so as to present significant results at the beginning and end. The difference between first and last tests p values suggests that authors do not present analyses randomly with respect to their statistical significance.

Significant variation in power among statistical tests was expected. For example, for a given sample size per group, the power of a one-sample t test is greater than that of a two-sample t test, because in the former case, the mean against which the data is compared is specified, whereas in the latter case, both means are estimated with error. Given the same underlying effect size being tested, this should result in lower p values for tests with greater statistical power. However, p values did not vary significantly among tests. This suggests either that the relationship is small and went undetected (power to detect a small effect was <36%) or that tests with greater power are used when testing for relationships with smaller actual effect sizes.

The first test per paper had significantly greater power and smaller p values than did the last test. This trend was consistent across journals, taxa, or statistical test type. In retrospect, the random selection of tests from papers might be desirable. This is, however, easier said than done, which was why we used the first and last test to remove any subjectivity on our part. In addition, despite these differences between first and last tests, both convey the same message: Power to detect small and medium effects is low.

Neither p values nor power varied significantly among the 10 journals (although $p < .10$ for p values of last tests and power of first tests). This implies that impact factor or any assessment of journal quality is unrelated to the statistical power or effect sizes they generally report. However, mean p values and power of first and last tests were significantly correlated across journals, which does suggest some degree of repeatability among journals.

Somewhat unexpectedly, neither p values nor statistical power varied significantly among the 11 taxa. (There was, however, about 10% less statistical power in studies of birds compared with insects for the last test per study.) In addition, neither mean p values nor power of first and last tests was correlated across taxa. We had initially assumed that sample sizes, hence power, would be larger for insects than, say, mammals. The lack of detectable variation may partly lie in the kinds of questions asked. For example, a study on primates may ask whether distance moved per day differed between two groups (a very specific question) based on 100 days of observations, whereas a study on insects might ask whether body size differed between mated and unmated males based on 100 males per group.

There was a significant negative correlation between mean power and *p* values. There are several possible explanations why smaller studies more often report nonsignificant results. First, true effect sizes may be smaller in study systems in which sample sizes tend to be smaller. We know of no reason why this should be true. Second, for a nonzero effect size, as sample size increases, power increases and *p* values decrease. So, if there is no underlying correlation between sample size and the true effect size, then one possible interpretation is that some tests fail to report a significant relationship because they lack statistical power. One could argue that there is good evidence for this conclusion because there should be considerable variability in the true effect sizes the 1362 tests were trying to detect. This will greatly reduce the strength of the pattern (e.g., in a study in which the true effect was large; even with a small sample size and low power, the reported *p* value will tend to be small). Third, researchers may adjust sample sizes based on their assessment of the likelihood of detecting an effect. For example, researchers may be disinclined to increase sample sizes when they infer that there is no significant effect to detect. This would also yield a negative correlation between power and *p* value. We believe there is a measure of truth to this because, for a given sample size, a researcher who finds that $p = .07$ is likely to continue collecting data in the hope of reaching $p < .05$, whereas a researcher faced with $p = .57$ is likely to conclude that she/he will not reach significance and therefore discontinue collecting data. This is a rational, but worrying, behavior because studies with significant results are more likely to be published than those without (Møller and Jennions, 2001; Palmer, 2000; Song et al., 2000).

Conclusions

Statistical power in behavioral ecology is distressingly low. The unavoidable solution is to increase sample sizes. It may be argued that logistic, ethical, conservation, and financial constraints make this impossible. It could be claimed with equal vigor that designing a study with low power is unethical and wasteful because nonsignificant findings are inconclusive. In many cases, especially when dealing with invertebrates, we suspect sample sizes can be increased. At present, many researchers decide on a sample size based on examination of previously published work (i.e., conventions among peers) rather than explicit consideration of power. If sample sizes are not readily increased, then it becomes even more important to conduct meta-analyses to detect broader trends (Cooper and Hedges, 1994). In medicine, meta-analysis of numerous small-scale studies (few of which are likely to detect significant trends) may provide a more cost-effective way of assessing the value of a new treatment than investing in a few large-scale studies (Song et al., 2000: 39). In addition, even in medicine, in which studies are on one species, there is the danger that a large study, no matter how well designed, may generate conclusions that can not be extrapolated to society at large if the study population is unrepresentative (e.g., smoking may greatly increase mortality in long-lived Western societies, but have little effect in developing countries where life expectancy is already low).

Biologists may have to show greater restraints when discussing the results of their own studies (when conclusions are hampered by low statistical power) and wait until sufficient studies have been conducted to determine general trends. Unfortunately, this is not how papers are currently written. A researcher whose modest conclusion is "more studies are needed until the results of my study can be interpreted" is probably less likely to be published than one who places a strong interpretation on his or her findings. This is true whether the results are significant or nonsignificant. Authors that extrapolate from a single significant study to the world at large commit an equal but opposite sin. The publication process needs to place greater emphasis on evaluation of the design and implementation of experiments or data collection protocols and less on the *p* values (or even the power) of the relationships detected (Palmer, 2000).

We believe there is a need to encourage the quantitative synthesis of the literature using modern meta-analytic techniques. Of course, as with any form of review, conclusions are only as reliable as the studies on which they are based. Those who are concerned that meta-analysis leads to "rubbish in, rubbish out" should be emphasizing the importance of a priori "quality" criteria for the inclusion of studies in meta-analyses. It is important to note that from a meta-analysis perspective, poor studies are not those with small sample sizes. These studies have little bearing on the outcome of a meta-analysis because effect sizes are weighted by their sampling variance, which is inversely related to sample size. In contrast, because narrative reviews do not explicitly consider the influence of sample size, they are far more likely to incorrectly estimate trends by giving equal weighting to studies that differ greatly in the extent to which we can trust their conclusions.

We thank Patricia Backwell and John Christy for advice and support. M.D.J. extends special thanks to Dr. Ira Rubinoff and the Smithsonian Tropical Research Institute for bridging funding during the course of this study. This work was partly funded by the Australian Research Council (grants A00104301 and F00104500).

REFERENCES

Bauchau V, 1997. Is there a "file drawer problem" in biological research? Oikos 79:407–409.

Bradley DR, Russell RL, Reeve CP, 1996. Statistical power in complex experimental designs. Behav Res Methods Instrum Comput 24:190–204.

Chase LJ, Chase LR, Tucker RK, 1978. Statistical power in physical anthropology: a technical report. Am J Phys Anthropol 49:133–137.

Chung KC, Kalliainen LK, Hayward RA, 1998. Type II (β) errors in the hand literature: the importance of power. J Hand Surg 23:20–25.

Cohen J, 1962. The statistical power of abnormal-social psychological research: a review. J Abnorm Soc Psychol 65:145–153.

Cohen J, 1988. Statistical power analysis for the behavioural sciences, 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum.

Cooper H, Hedges LV, eds, 1994. The handbook of research synthesis. New York: Russell Sage Foundation.

Csada RD, James PC, Espie RHM, 1996. The "file drawer problem" of non-significant results: does it apply to biological research? Oikos 76:591–593.

Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I, 2000. Size and quality of randomised controlled trials in head injury: review of published studies. Br Med J 320:1308–1311.

Erdfelder E, Faul F, Buchner A, 1996. G*power: a general power analysis program. Behav Res Methods Instrum Comput 28:1–11.

Greenwood J JD, 1993. Statistical power. Anim Behav 46:1011.

Hoenig JM, Heisey DM, 2001. The abuse of power: the pervasive fallacy of power calculation for data analysis. Am Stat 55:19–24.

Hurlbert SH, 1994. Old shibboleths and new syntheses. Trends Ecol Evol 9:495–496.

Jennions MD, Møller AP, 2002a. Publication bias in ecology and evolution: an empirical assessment using the "trim and fill" method. Biol Rev 77:211–222.

Jennions MD, Møller AP, 2002b. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proc R Soc Lond Ser B 269:43–48.

Kazantzis N, 2000. Power to detect homework effects in psychotherapy outcome research. J Consult Clin Psych 68:166–170.

Kloster KL, Layne BH, 1997. Low power, Type II errors, and other statistical problems in recent cardiovascular research. Am J Physiol 273:487–493.

Kotiaho JS, Tomkins JL, 2002. Meta-analysis can it ever fail? Oikos 96:551–553.

Maddock JE, Rossi JS, 2001. Statistical power of articles published in three health psychology-related journals. Health Psychol 20:76–78.

Mengel MB, Davis AB, 1993. The statistical power of family practice research. Family Pract Res 13:105–111.

Moher D, Dulberg CS, Wells GA, 1994. Statistical power, sample size, and their reporting in randomized controlled trials. J Am Med Assoc 272:122–124.

Møller AP, Jennions MD, 2001. Testing and adjusting for publication bias. Trends Ecol Evol 16:580–586.

Møller AP, Jennions MD, 2002. How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132:492–500.

Noor MA, Smith KR, 2000. Recombination, statistical power, and genetic studies of sexual isolation in *Drosophila*. J Hered 91:99–103.

Palmer AR, 1999. Detecting publication bias in meta-analyses: a case study of fluctuating asymmetry and sexual selection. Am Nat 154:220–233.

Palmer AR, 2000. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. Ann Rev Ecol Syst 31:441–480.

Peres-Neto PR, Olden JD, 2001. Assessing the robustness of randomization tests: examples from behavioural studies. Anim Behav 61:79–86.

Peterman RM, 1990. Statistical power analysis can improve fisheries research and management. Can J Fish Aquatic Sci 47:2–15.

Polit DF, Sherman RE, 1990. Statistical power in nursing research. Nursing Res 39:365–369.

Reeve HK, 2001. In search of unified theories in sociobiology: help from social wasps. In: Model systems in behavioral ecology (Dugatkin LA, ed). Princeton, New Jersey: Princeton University Press; 57–71

Rosenberg MS, Adams DC, Gurevitch J, 2000. MetaWin: statistical software for meta-analysis, version 2.0. Sunderland, Massachusetts: Sinauer Associates.

Rossi JS, 1990. Statistical power of psychological research: what have we gained in 20 years? J Consult Clin Psychol 58:646–656.

Sedlmeier P, Gigerenzer G, 1989. Do studies of statistical power have an effect on the power of studies? Psychol Bull 105:309–316.

Siegel S, Castellan NJ Jr, 1988. Nonparametric statistics for the behavioural sciences, 2nd ed. Singapore: McGraw-Hill.

Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ, 2000. Publication and related biases. Health Technol Assess 4(10):1–115.

Stoehr AM, 1999. Are significance thresholds appropriate for the study of animal behaviour? Anim Behav 57:F22–F25.

Thomas L, Juanes F, 1996. The importance of statistical power analysis: an example from *Animal Behaviour*. Anim Behav 52: 856–859.

Thomas L, Krebs CJ, 1997. A review of statistical power analysis software. Bull Ecol Soc Amer 78:126–139.

Thompson CF, Neill AJ, 1993. Statistical power and accepting the null hypothesis. Anim Behav 46:1012.

Toft CA, Shea PJ, 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. Am Nat 122: 618–625.

Zar JH, 1999. Biostatistical analysis, 4th ed. London: Prentice-Hall.