The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis

John M. HOENIG and Dennis M. HEISEY

It is well known that statistical power calculations can be valuable in planning an experiment. There is also a large literature advocating that power calculations be made whenever one performs a statistical test of a hypothesis and one obtains a statistically nonsignificant result. Advocates of such post-experiment power calculations claim the calculations should be used to aid in the interpretation of the experimental results. This approach, which appears in various forms, is fundamentally flawed. We document that the problem is extensive and present arguments to demonstrate the flaw in the logic.

KEY WORDS: Bioequivalence testing; Burden of proof; Observed power; Retrospective power analysis; Statistical power; Type II error.

1. INTRODUCTION

It is well known among applied scientists that a lack of impact or effect is not sufficiently established by a failure to demonstrate statistical significance. A failure to reject the null hypothesis of no effect may be the result of low statistical power when an important effect actually exists and the null hypothesis of no effect is in fact false. This can be called the dilemma of the nonrejected null hypothesis: what should we do when we fail to reject a hypothesis?

Dismayingly, there is a large, current literature that advocates the inappropriate use of post-experiment power calculations as a guide to interpreting tests with statistically nonsignificant results. These ideas are held tenaciously in a variety of disciplines as evidenced by methodological recommendations in 19 applied journals (Table 1). In our experience as consulting statisticians, authors are not infrequently required to perform such calculations by journal reviewers or editors; at least two journals ask for these calculations as a matter of policy (Anon. 1995; Anon. 1998). We emphasize that these calculations are sought primarily with the thought that they are useful for explaining the observed data, rather than for the purpose of planning some future experiment. We even found statistical textbooks that illustrate the flawed approach (e.g., Rosner 1990; Winer, Brown, and Michels 1991; Zar 1996). Researchers need to be made aware of the shortcomings of power calculations as data analytic tools and taught more appropriate methodology.

It is important to understand the motivation of applied scientists for using power analysis to interpret hypothesis tests with nonsignificant results. The traditional, widely accepted standard has been to protect the investigator from falsely concluding that some treatment has an effect when indeed it has none. However, there is increasing recognition that a "reversal of the usual scientific burden of proof" (e.g., Dayton 1998) is preferred in many areas of scientific inference. Areas where this is a particular concern include making decisions about environmental impacts, product safety, and public welfare where some people want to be protected from failing to reject a null hypothesis of no impact when a serious (e.g., harmful or dangerous) effect exists. We believe that the post-hoc power approaches that have consequently arisen are due to applied scientists being heavily tradition-bound to test the usual "no impact null hypothesis," despite it not always being the relevant null hypothesis for the question at hand.

We describe the flaws in trying to use power calculations for data-analytic purposes and suggest that statistics courses should have more emphasis on the investigator's choice of hypotheses and on the interpretation of confidence intervals. We also suggest that introducing the concept of equivalence testing may help students understand hypothesis tests. For pedagogical reasons, we have kept our explanations as simple as possible.

2. INAPPROPRIATE USES OF POWER ANALYSIS

2.1 "Observed Power"

There are two common applications of power analysis when a nonrejected null hypothesis occurs. The first is to compute the power of the test for the observed value of the test statistic. That is, assuming the observed treatment effects and variability are equal to the true parameter values, the probability of rejecting the null hypothesis is computed. This is sometimes referred to as "observed power." Several widely distributed statistical software packages, such as SPSS, provide observed power in conjunction with data analyses (see Thomas and Krebs 1997). Advocates of observed power argue that there is evidence for the null hypothesis being true if statistical significance was not achieved despite the computed power being high at the observed effect size. (Usually, this is stated in terms of the evidence for the null hypothesis (no effect) being weak if observed power was low.)

John M. Hoenig is Professor, Virginia Institute of Marine Science, College of William and Mary, Gloucester Point, VA 23062 (E-mail: hoenig@vims.edu). Dennis M. Heisey is Statistician, Department of Surgery and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53792. Order of authorship determined by randomization. The authors thank Marilyn Lewis for research assistance and the anonymous reviewers for helpful comments. This is VIMS Contribution No. 2335.

Table 1. Journals with Articles Advocating Post-Experiment Power Analysis

American Journal of Physical Anthropology: Hodges and Schell (1988)
American Naturalist: Toft and Shea (1983); Rotenberry and Wiens (1985)
**Animal Behavior*: Thomas and Juanes (1996); Anon. (1998)
Aquaculture: Searcy-Bernal (1994)
Australian Journal of Marine and Freshwater Research: Fairweather (1991)
Behavioral Research Therapy: Hallahan and Rosenthal (1996)
Bulletin of the Ecological Society of America: Thomas and Krebs (1997)
Canadian Journal of Fisheries and Aquatic Sciences: Peterman (1989, 1990a)
Conservation Biology: Reed and Blaustein (1995, 1997); Hayes and Steidl (1997); Thomas (1997)
Ecology: Peterman (1990b)
Journal of Counseling Psychology: Fagley (1985)
**Journal of Wildlife Management*: (Anon., 1995); Steidl, Hayes and Schauber (1997)
Marine Pollution Bulletin: Peterman and M'Gonigle (1992)
Neurotoxicology and Teratology: Muller and Benignus (1992)
Rehabilitation Psychology: McAweeney, Forchheimer, and Tate (1997)
Research in the Teaching of English: Daly and Hexamer (1983)
Science: Dayton (1998)
The Compass of Sigma Gamma Epsilon: Smith and Kuhnhenn (1983)
Veterinary Surgery: Markel (1991)

NOTE: $*$ indicates journal requires or requests post-experiment power calculations when test results are nonsignificant.

Observed power can never fulfill the goals of its advocates because the observed significance level of a test ("p value") also determines the observed power; for any test the observed power is a 1:1 function of the p value. A p value is a random variable, P, on $[0, 1]$. We represent the cumulative distribution function (cdf) of the p value as $\Pr(P \leq p) = G_\delta(p)$, where δ is the parameter value. Consider a one-sample Z test of the hypothesis $\mathrm{H}_0 : \mu \leq 0$ versus $\mathrm{H}_a : \mu > 0$ when the data are from a normal distribution with known σ. Let $\delta = \sqrt{n}\mu/\sigma$. Then $G_\delta(p) = 1-\Phi(Z_p-\delta)$, where Z_p is the $100(1-p)$th percentile of the standard normal distribution (Hung, O'Neill, Bauer, and Kohne 1997). That is, Z_p is the observed statistic. Both p values and observed power are obtained from $G_\delta(p)$. A p value is obtained by setting $\mu = 0$, so $G_0(p) = 1 - \Phi(Z_p) = p$. Observed power is obtained by setting the parameter to the observed statistic and finding the percentile for $P < \alpha$, so observed power is given by $G_{Z_p}(\alpha) = 1 - \Phi(Z_\alpha - Z_p)$ and thus the observed power is determined completely by the p value and therefore adds nothing to the interpretation of results. An interesting special case occurs when $P = \alpha$; for the Z test example above it is immediately obvious that observed power $= .5$ because $Z_\alpha = Z_p$. Thus, computing observed power can never lead to a statement such as "because the null hypothesis could not be rejected and the observed power was high, the data support the null hypothesis." Because of the one-to-one relationship between p values and observed power, nonsignificant p values always correspond to low observed powers (Figure 1). Computing the observed power after observing the p value should cause nothing to change about our interpretation of the p value. These results are easily extended to two-sided tests.

There is a misconception about the relationship between observed power and p value in the applied literature which is likely to confuse nonstatisticians. Goodman and Berlin (1994), Steidl, Hayes, and Schauber (1997), Hayes and Steidl (1997), and Reed and Blaustein (1997) asserted without proof that observed power will *always* be less than .5 when the test result is nonsignificant. An intuitive counterexample is as follows. In a two-tailed Z test, the test statistic has the value $Z = 1.96$ if the test is marginally significant at $\alpha = .05$. Therefore, the probability of observing a test statistic above 1.96, if the true mean of Z is 1.96, is .5. The probability of rejecting the null hypothesis is the probability of getting a test statistic above 1.96 or below -1.96. Therefore, the probability is slightly *larger* than .5. In fact, it is rather easy to produce special examples of test statistics with skewed distributions that can produce arbitrarily high observed powers for $p = \alpha$.

A number of authors have noted that observed power may not be especially useful, but to our knowledge a fatal logical flaw has gone largely unnoticed. Consider two experiments that gave rise to nonrejected null hypotheses. Suppose the observed power was larger in the first experiment than the second. Advocates of observed power would interpret this to mean that the first experiment gives stronger support *favoring* the null hypothesis. Their logic is that if power is

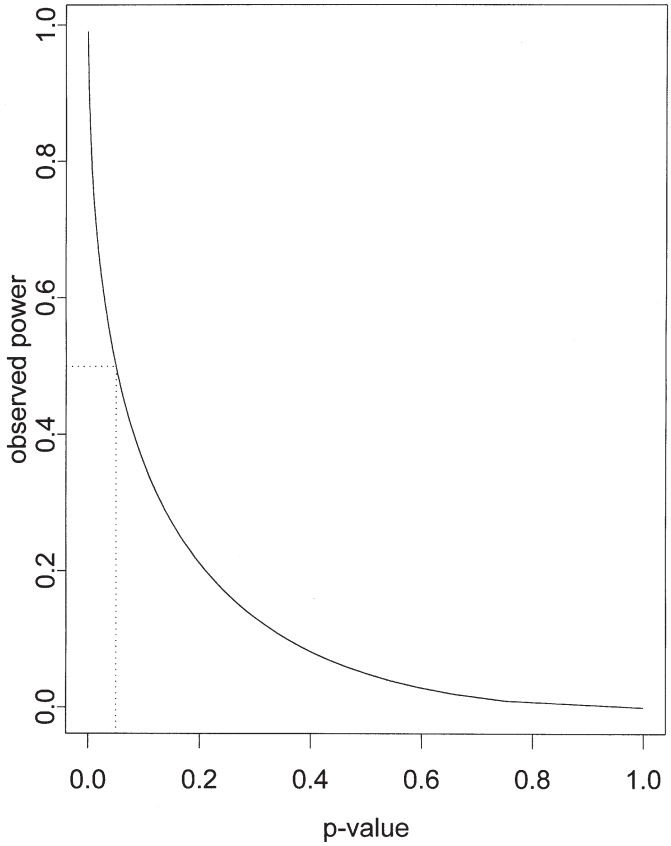

Figure 1. "Observed" Power as a Function of the p Value for a One-Tailed Z Test in Which α is Set to .05. When a test is marginally significant (P = .05) the estimated power is 50%.

low one might have missed detecting a real departure from the null hypothesis but if, despite high power, one fails to reject the null hypothesis, then the null is probably true or close to true. This is easily shown to be nonsense. For example, consider the one-sided Z test described above. Let Z_{p_1} and Z_{p_2} refer to the observed test statistics in the respective experiments. The observed power was highest in the first experiment and we know this implies $Z_{p_1} > Z_{p_2}$ because observed power is $G_{Z_p}(\alpha)$ which is an increasing function of the Z statistic. So by usual standards of using the p value as statistical evidence, the first experiment gives the stronger support *against* the null, contradicting the power interpretation. We will refer to this inappropriate interpretation as the "power approach paradox" (PAP): higher observed power does not imply stronger evidence for a null hypothesis that is not rejected.

2.2 "Detectable Effect Size" and "Biologically Significant Effect Size"

A second, perhaps more intriguing, application of post-experiment power calculations is finding the hypothetical true difference that would have resulted in a particular power, say .9. This is an attempt to determine the "detectable effect size." It is applied as follows: an experiment is performed that fails to reject the null. Then, based on the observed variability, one computes what the effect size would have needed to have been to have a power of .9. Advocates of this approach view this "detectable effect size" as an upper bound on the true effect size; that is, because significance was not achieved, nature is unlikely to be near this state where power is high. The closer the detectable effect size is to the null hypothesis of 0, the stronger the evidence is taken to be for the null. For example, in a one-tailed Z test of the hypothesis $\mathrm{H}_0 : \mu \leq 0$ versus $\mathrm{H}_a : \mu > 0$, one might observe a sample mean $\overline{X} = 1.4$ with $\sigma_{\overline{X}} = 1$. Thus, $Z = 1.4$ and $P = .08$, which is not significant at $\alpha = .05$. We note that if the true value of μ were 3.29 (and $\sigma_{\overline{X}}$ were 1) we would have power $= .95$ to reject H_0. Hence, 3.29 would be considered an upper bound on the likely value of the true mean. (Note that a 95% upper confidence bound on μ would be 3.04. We return to this point later.)

A variant of the "detectable effect size" approach is the "biologically significant effect size" approach, where one computes the power at some effect size deemed to be biologically important. The higher the computed power is for detecting meaningful departures from the null, the stronger the evidence is taken to be for nature to be near the null when the null is not rejected.

These inferential approaches have not been justified formally. Cohen (1988, p. 16) claimed that if you design a study to have high power $1 - \beta$ to detect departure Δ from the null hypothesis, and you fail to reject the null hypothesis, then the conclusion that the true parameter value lies within Δ units of the null value is "significant at the β level. Thus, in using the same logic as that with which we reject the null hypothesis with risk equal to α, the null hypothesis can be accepted in preference to that which holds that ES [the effect size] $= \Delta$ with risk equal to β." (We have changed Cohen's notation in the above to conform to that used here.) Furthermore, Cohen stated (p. 16) "'proof' by statistical induction is probabilistic" without elaboration. He appeared to be making a probabilistic statement about the true value of the parameter which is invalid in a classical statistical context. Furthermore, because his procedure chooses the sample size to have a specified, fixed power before conducting the experiment, his argument assumes that the actual power is equal to the intended power and, additionally, his procedure ignores the experimental evidence about effect size and sampling variability because the value of β is not updated according to the experimental results. Rotenberry and Wiens (1985) and Searcy-Bernal (1994) cited Cohen in justifying their interpretation of post-experiment computed power.

Although many find the detectable effect size and biologically significant effect size approaches more appealing than the observed power approach, these approaches also suffer from fatal PAP. Consider the previous two experiments where the first was closer to significance; that is, $Z_{p_1} > Z_{p_2}$. Furthermore, suppose that we observed the same estimated effect size in both experiments and the sample sizes were the same in both. This implies $\sigma_1 < \sigma_2$. For some desired level of power Π_α, one solves $\Pi_\alpha = 1 - \Phi(Z_\alpha - \sqrt{n}\rho/\sigma)$ for ρ to obtain the desired detectable effect size, ρ. It follows that the computed detectable effect size will be smaller in the first experiment. And, for any conjectured effect size, the computed power will always be higher in the first experiment. These results lead to the nonsensical conclusion that the first experiment provides the stronger evidence for the null hypothesis (because the apparent power is higher but significant results were not obtained), in direct contradiction to the standard interpretation of the experimental results (p values).

Various suggestions have been made for "improving" post-experiment power analyses. Some have noted certain estimates of general effect sizes (e.g., noncentrality parameters) may be biased (Thomas 1997; Gerard, Smith, and Weerakkody 1998), which potentially could be corrected. Others have addressed the fact that the standard error used in power calculations is known imprecisely, and have suggested computing confidence intervals for post-experiment power estimates (Thomas 1997; Thomas and Krebs 1997). This is curious because, in order to evaluate a test result, one apparently needs to examine power but, in order to evaluate (test) if power is adequate one does not consider the power of a test for adequate power. Rather, one switches the inferential framework to one based on confidence intervals. These suggestions are superfluous in that they do nothing to correct the fundamental PAP.

3. POWER ANALYSIS VERSUS CONFIDENCE INTERVALS

From a pedagogic point of view, it is interesting to compare the inference one would obtain from consideration of confidence intervals to that obtained from the power analysis approach. Confidence intervals have at least two interpretations. One interpretation is based on the equivalence

of confidence intervals and hypothesis tests. That is, if a confidence interval does not cover a hypothesized parameter value, then the value is refuted by the observed data. Conversely, all values covered by the confidence interval could not be rejected; we refer to these as the set of nonrefuted values. If the nonrefuted states are clustered tightly about a specific null value, one has confidence that nature is near the null value. If the nonrefuted states range widely from the null, one must obviously be cautious about interpreting the nonrejection as an indication of a "near-null" state. The more widely known interpretation is that confidence intervals cover the true value with some fixed level of probability. Using either interpretation, the breadth of the interval tells us how confident we can be of the true state of nature being close to the null.

Once we have constructed a confidence interval, power calculations yield no additional insights. It is pointless to perform power calculations for hypotheses outside of the confidence interval because the data have already told us that these are unlikely values. What about values inside the confidence interval? We already know that these are values that are not refuted by the data. It would be a mistake to conclude that the data refute any value within the confidence interval. However, there can be values within a 95% confidence interval that yield computed powers of nearly .975. Thus, it would be a mistake to interpret a value associated with high power as representing some type of upper bound on the plausible size of the true effect, at least in any straightforward sense. The proposition that computed power for effect sizes within a confidence interval can be very high can be demonstrated as follows. Consider the case where the random variable X has a normal distribution. We wish to test the null hypothesis that the mean is zero versus the alternative that it is not zero. A random sample of large size is taken which has a mean, $\overline{x}$, of 2 and a standard error of the mean of 1.0255. The upper critical region for a two-sided Z test then corresponds to values of the mean greater than $1.96 \times 1.0255 = 2.01$. Therefore, we fail to reject the null hypothesis. A 95% confidence interval would be $(-.01, 4.01)$. We note that a value of 4 for the population mean is not refuted by the data. Now post-hoc power calculation indicates the probability of rejecting the null hypothesis if the mean is actually 4 is $\Pr(|\overline{X}| > 2.01) = \Pr(Z > (2.01 - 4)/1.0255) + \Pr(Z < (-2.01 - 4)/1.0255)$ which is about .974. Thus, the power calculation suggests that a value of 4 for the mean is unlikely—otherwise we ought to have rejected the null hypothesis. This contradicts the standard theory of hypothesis tests.

4. EQUIVALENCE TESTING

Simply saying that an experiment demonstrates that a treatment is "near-null" because the confidence interval is narrow about the null value may seem unsatisfactorily "seat-of-the-pants." However, this can be formulated as a rigorous test. Suppose that we are willing to conclude that a treatment is negligible if its absolute effect is no greater than some small positive value Δ. Demonstrating such practical equivalence requires reversing the traditional burden of proof; it is not sufficient to simply fail to show a difference, one must be fairly certain that a large difference does not exist. Thus, in contrast to the traditional casting of the null hypothesis, the null hypothesis becomes that a treatment has a large effect, or $\mathrm{H}_0 : |D| \geq \Delta$, where D is the actual treatment effect. The alternative hypothesis is the hypothesis of practical equivalence, or $\mathrm{H}_A : |D| < \Delta$.

Schuirmann (1987) showed that if a $1 - 2\alpha$ confidence interval lies entirely between $-\Delta$ and Δ, we can reject the null hypothesis of nonequivalence in favor of equivalence at the α level. The equivalence test is at the α level because it involves two one-tailed α level tests, which together describe a $1 - 2\alpha$ level confidence interval. This approach to equivalence testing is actually always a bit on the conservative side; the actual level α' for normally distributed data from a one-sample experiment with known σ and nominal level α is $\alpha' = \alpha - 1 + \Phi\left(2\Delta\sqrt{n}/\sigma - Z_\alpha\right)$, which shows the conservatism will be slight in many practical applications where $2\Delta\sqrt{n}/\sigma$ substantially exceeds Z_α. More powerful equivalence testing procedures exist (e.g., Berger and Hsu 1996), but for well-behaved problems with simple structures the simplicity of this approach seems to make it a compelling choice to recommend to the researcher involved in analysis (Hauck and Anderson 1996).

Considering the power approach as a formal test in the above equivalence testing framework makes it clear why it is logically doomed. The power approach requires two outcomes before declaring equivalence, which are (1) the null hypothesis of no difference $\mathrm{H}_0 : D = \mathbf{0}$ cannot be rejected, and (2) some predetermined level of power must be achieved for $|D| = \Delta$. To achieve outcome 1, the absolute value of the observed test statistic must be less than Z_α. This in turn implies that the observed absolute difference $|d|$ must be less than $Z_a\sigma/\sqrt{n}$. Thus, as $|D|$ becomes more precisely estimated by increasing n or decreasing σ, the observed difference $|d|$ must become progressively smaller if we want to demonstrate equivalence. This simply does not make sense: it should become easier, not more difficult, to conclude equivalence as $|D|$ becomes better characterized. Schuirmann (1987) noted that when viewed as a formal test of equivalence, the power approach results in a critical region that is essentially upside down from what a reasonable equivalence test should have.

5. DISCUSSION

Because of the prominence of post-hoc power calculations for data analysis in the literature, elementary statistics texts should devote some attention to explaining what should not be done. However, there is a larger lesson to be learned from the confusion about power analysis. We believe the central focus of good data analysis should be to find which parameter values are supported by the data and which are not. Perhaps unwittingly, advocates of post-hoc power analysis are seemingly grappling with exactly this question.

The reader with Bayesian inclinations would probably think "what foolishness—the whole issue would be moot if people just focused on the sensible task of obtaining poste-

rior distributions." Philosophically, we find this attractive as it avoids some nagging issues in frequentist statistics concerning p values and confidence intervals (e.g., Berry 1993; Freeman 1993; Schervish 1996; Goodman 1999a,b). But, the real world of data analysis is for the most part solidly frequentist and will remain so into the foreseeable future. Within the limitations of the frequentist framework, it is important that analyses be as appropriate as possible.

Introductory statistics classes can focus on characterizing which parameter values are supported by the data by emphasizing confidence intervals more and placing less emphasis on hypothesis testing. One might argue that a rigorous understanding of confidence intervals requires a rigorous understanding of hypothesis testing and p values. We feel that researchers often do not need a rigorous understanding of confidence intervals to use them to good advantage. Although we cannot demonstrate it formally, we suspect that imperfectly understood confidence intervals are more useful and less dangerous than imperfectly understood p values and hypothesis tests. For example, it is surely prevalent that researchers interpret confidence intervals as if they were Bayesian credibility regions; to what extent does this lead to serious practical problems? The indirect logic of frequentist hypothesis testing is simply nonintuitive and hard for most people to understand (Berry 1993; Freeman 1993; Goodman 1999a,b). If informally motivated confidence intervals lead to better science than rigorously motivated hypothesis testing, then perhaps the rigor normally presented to students destined to be applied researchers can be sacrificed.

Of course, researchers must be exposed to hypothesis tests and p values in their statistical education if for no other reason than so they are able to read their literatures. However, more emphasis should be placed on general principles and less emphasis on mechanics. Typically, almost no attention is given to why a particular null hypothesis is chosen and there is virtually no consideration of other options. As Hauck and Anderson (1996) noted, both statisticians and nonstatisticians often test the wrong hypothesis because they are so conditioned to test null hypotheses of no difference. Statisticians need to be careful not to present statistical analysis as a rote process. Introductory statistics students frequently ask the question, "why focus on protection against erroneously rejecting a true null of no difference?" The stock answer is often something like "it is bad for science to conclude a difference exists when it does not." This is not sufficient. In matters of public health and regulation, it is often more important to be protected against erroneously concluding no difference exists when one does. In any particular analysis, one needs to ask whether it is more appropriate to use the no difference null hypothesis rather than the nonequivalence null hypothesis. This is a question that regulators, researchers, and statisticians need to be asked and be asking constantly. We doubt whether many researchers are even aware that they have choices with respect to the null hypotheses they test and that the choices reflect where the burden of proof is placed.

We would not entirely rule out the use of power-type concepts in data analysis, but their application is extremely limited. One potential application might be to examine whether several experiments were similar, except for sample size; this might be an issue for example in meta-analyses (Hung, O'Neill, Bauer, and Kohne 1997). The goal here, examining homogeneity, differs from the usual motivations for post hoc power considerations.

Power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature. With traditional frequentist statistics, this is best achieved with confidence intervals, appropriate choices of null hypotheses, and equivalence testing. Confusion about these issues could be reduced if introductory statistics classes for researchers placed more emphasis on these concepts and less emphasis on hypothesis testing.

[Received July 2000. Revised September 2000.]

REFERENCES

Anon. (1995), "Journal News," *Journal of Wildlife Management*, 59, 196–199.

——— (1998), "Instructions to Authors," *Animal Behavior*, 55, i–viii.

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," *Statistical Science*, 11, 283–319.

Berry, D. A. (1993), "A Case for Bayesianism in Clinical Trials," *Statistics in Medicine*, 12, 1377–1393.

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

Daly, J. A., and Hexamer, A. (1983), "Statistical Power in Research in English Education," *Research in the Teaching of English*, 17, 157–164.

Dayton, P. K. (1998), "Reversal of the Burden of Proof in Fisheries Management," *Science*, 279, 821–822.

Fagley, N. S. (1985), "Applied Statistical Power Analysis and the Interpretation of Nonsignificant Results by Research Consumers," *Journal of Counseling Psychology*, 32, 391–396.

Fairweather, P. G. (1991), "Statistical Power and Design Requirements for Environmental Monitoring," *Australian Journal of Marine and Freshwater Research*, 42, 555–567.

Freeman, P. R. (1993), "The Role of P values in Analysing Trial Results," *Statistics in Medicine*, 12, 1443–1452.

Gerard, P. D., Smith, D. R., and Weerakkody, G. (1998), "Limits of Retrospective Power Analysis," *Journal of Wildlife Management*, 62, 801–807.

Goodman, S. N. (1999a), "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

——— (1999b), "Toward Evidence-Based Medical Statistics 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013.

Goodman, S. N., and Berlin, J. A. (1994), "The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results," *Annals of Internal Medicine*, 121, 200–206.

Hallahan, M., and Rosenthal, R. (1996), "Statistical Power: Concepts, Procedures, and Applications," *Behavioral Research Therapy*, 34, 489–499.

Hayes, J. P., and Steidl, R. J. (1997), "Statistical Power Analysis and Amphibian Population Trends," *Conservation Biology*, 11, 273–275.

Hauck, W. W., and Anderson, S. (1996), Comment on "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," *Statistical Science*, 11, 303–304.

Hodges, D. C., and Schell, L. M. (1988), "Power Analysis in Biological Anthropology," *American Journal of Physical Anthropology*, 77, 175–181.

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997), "The Behavior of the P Value When the Alternative Hypothesis is True," *Biometrics*, 53, 11–22.

Markel, M. D. (1991), "The Power of a Statistical Test—What Does Insignificance Mean?," *Veterinary Surgery*, 20, 209–214.

McAweeney, M. J., Forchheimer, M., and Tate, D. G. (1997), "Improving Outcome Research in Rehabilitation Psychology: Some Methodological Recommendations," *Rehabilitation Psychology*, 42, 125–135.

Muller, K. E., and Benignus, V. A. (1992), "Increasing Scientific Power With Statistical Power," *Neurotoxicology and Teratology*, 14, 211–219.

Peterman, R. (1989), "Application of Statistical Power Analysis to the Oregon Coho Salmon (*Oncorhynchus kisutch*) Problem," *Canadian Journal of Fisheries and Aquatic Sciences*, 46, 1183.

——— (1990a), "Statistical Power Analysis Can Improve Fisheries Research and Management," *Canadian Journal of Fisheries and Aquatic Sciences*, 47, 2–15.

——— (1990b), "The Importance of Reporting Statistical Power: the Forest Decline and Acidic Deposition Example," *Ecology*, 71, 2024–2027.

Peterman, R., and M'Gonigle, M. (1992), "Statistical Power Analysis and the Precautionary Principle," *Marine Pollution Bulletin*, 24, 231–234.

Reed, J. M., and Blaustein, A. R. (1995), "Assessment of 'Nondeclining' Amphibian Populations Using Power Analysis," *Conservation Biology*, 9, 1299–1300.

——— (1997), "Biologically Significant Population Declines and Statistical Power," *Conservation Biology*, 11, 281–282.

Rosner, B. (1990), *Fundamentals of Biostatistics* (3rd ed.), Boston: PWS-Kent Publishing.

Rotenberry, J. T., and Wiens, J. A. (1985), "Statistical Power Analysis and Community-Wide Patterns," *American Naturalist*, 125, 164–168.

Schervish, M. J. (1996), "P Values: What They Are and What They Are Not," *Journal of the American Statistical Association*, 50, 203–206.

Schuirmann, D. J. (1987), "A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657–680.

Searcy-Bernal, R. (1994), "Statistical Power and Aquacultural Research," *Aquaculture*, 127, 371–388.

Smith, A. D., and Kuhnhenn, G. L. (1983), "A Statistical Note on Power Analysis as Applied to Hypothesis Testing Among Selected Petrographic Point-Count Data," *The Compass of Sigma Gamma Epsilon*, 61, 22–30.

Steidl, R. J., Hayes, J. P., and Schauber, E. (1997), "Statistical Power Analysis in Wildlife Research," *Journal of Wildlife Management*, 61, 270–279.

Thomas, L. (1997), "Retrospective Power Analysis," *Conservation Biology*, 11, 276–280.

Thomas, L., and Juanes, F. (1996), "The Importance of Statistical Power Analysis: An Example from Animal Behaviour," *Animal Behavior*, 52, 856–859.

Thomas, L., and Krebs, C. J. (1997), "A Review of Statistical Power Analysis Software," *Bulletin of the Ecological Society of America*, 78, 126–139.

Toft, C. A., and Shea, P. J. (1983), "Detecting Community-wide Patterns: Estimating Power Strengthens Statistical Inference," *American Naturalist*, 122, 618–625.

Winer, B. J., Brown, D. R., and Michels, K. M. (1991), *Statistical Principles in Experimental Design* (3rd ed.), New York: McGraw-Hill.

Zar, J. H. (1996), *Biostatistical Analysis* (3rd ed.), Upper Saddle River, NJ: Prentice Hall.