Tutorials in Quantitative Methods for Psychology
2007, Vol. 3 (2), p. 28- 34. DOI: 10.20982/tqmp.03.2.p028

Statistical power: An historical introduction

Jean Descôteaux
Université de Sherbrooke

Despite the funding agencies' growing demands towards power analyses, we believe researchers are still not fully aware of the statistical power concept, of the possible benefits of power analysis in the planning phase and of the ways to increase the chances of significantly detecting a given effect in their study. The following review falls within this area of interest. We discuss the history of the concept of statistical power, the reasons for its ongoing neglect, its potential benefits to researchers, as well as actual ways to improve statistical power. We also touch upon the subject of the impact of power analysis on the scientific literature.

The concept of statistical power is not new. It was formulated in the 1930's by Jerzy Neyman, a Moldavian who later immigrated to the United States, and Egon S. Pearson, the son of Karl Pearson, the British statistician who introduced the famous *r* (Neyman & Pearson, 1928, 1933). While considered promising, the concept, however, never gained much popularity. Its application in scientific research planning was strongly opposed by Sir Ronald Fisher (also from Great Britain), an influential figure in the field of statistics at the time, which might explain why statistical power remained relatively unknown until Jacob Cohen (USA) brought it back to light in the early 1960's. Interest in statistical power was revived partially thanks to Cohen's article (1962) in which he described the insufficient statistical power of studies published in the 1960 volume of the *Journal of Abnormal and Social Psychology*. In his review, Cohen concluded that the reported studies had, on average, a less than 1 in 2 probability (a 48% chance) to obtain statistical confirmation of an actual medium effect.

Since the publication of Cohen's analysis, the concept of statistical power has gained ground, evidenced by the sharp increase in the number of references to the popular book by Cohen (1969, 1977) in the scientific literature from 4 instances in 1971 to 214 instances in 1987 (Sedlmeier & Gigerenzer, 1989). However, despite this apparent growing awareness of the concept, the inadequacy in terms of insufficient statistical power described by Cohen in 1962 still exists to date. Rossi (1990) applied the method used by Cohen in 1962 to several reports published in 1982 in the *Journal of Abnormal Psychology*, the *Journal of Consulting and Clinical Psychology* and the *Journal of Personality and Social Psychology* and found that these studies had, on average, at most a 57% probability to obtain statistical confirmation of a medium effect, which is comparable to the 48% reported by Cohen in 1962. For their part, Sedlmeier and Gigerenzer (1989) used the same method to analyze reports published in 1984 in the *Journal of Abnormal Psychology* and showed that these studies had, on average, no more than a 37% chance of finding an actual medium effect. More recently, Bezeau and Graves (2001), Clark-Carter (1997), Kosciulek and Szymanski (1993), and Mone, Mueller, and Mauland (1996) reported a similar lack of power in work published in such diverse areas as clinical neuropsychology, articles reported in *British Journal of Psychology*, rehabilitation counseling research, and management. The only exception to the rule was provided by Maddock and Rossi (2001), who showed that research in three health-related journals (*Health Psychology*, *Addictive Behaviors*, and *Journal of Studies on Alcohol*) published in 1997 had adequate power to detect large and medium effects.

In general, the latest results speak of a profound paradox. In fact, it is hard to explain why, despite the great importance that reviewers attribute to the significance of test results and despite the ever growing difficulty to obtain research funding, research workers seem content with experimental protocols that yield inconclusive results in 1 out of 2 cases. Would they rather spend time and money to little avail than actually plan their research to ensure sufficient statistical power (a power of .80, for example)? Admittedly, the concept has gained ground. The American

Psychological Association does its part in popularizing the idea, urging researchers, for instance, to include at least some index of effect size in their results section (APA, 2001). Journals such as the *Journal of Clinical and Consulting Psychology* also contribute by specifically instructing authors to report effect sizes for primary study findings as well as confidence intervals for them (Instructions to Authors, *Journal of Consulting and Clinical Psychology*, 2007). So, if today's researchers are more alert to statistical power, it seems they are still unconvinced about its possible benefits in the planning phase and are not fully aware of the ways to increase the chances of significantly detecting a given effect in their study. The following review falls within this area of interest. We discuss the concept of statistical power, the reasons for its ongoing neglect, its potential benefits to researchers, as well as actual ways to improve statistical power. We also touch upon the subject of the impact of power analysis on the scientific literature.

Statistical power, definition and application

Simply put, the power of a statistical test is the probability that the test will yield statistically significant results, given the existence of an actual effect. While seemingly simple, this definition encompasses a mathematical complexity that often discourages the uninitiated. Statistical power is determined by various criteria, such as the sample size (N), the effect size of the observed phenomenon (e.g. d) and the applied level of statistical significance (α). The mathematical relationship between the four elements allows having any of these parameters quantified as a function of the three others (Cohen, 1988).

While adding a bit of complexity, such interrelations allow for flexibility in the application of the statistical power concept. For example, the calculation of power as a function of the other parameters is particularly useful in the research-planning phase (*a priori*) or to quantify the power of a completed test (*a posteriori*). One of the most commonly used applications of the statistical power concept is to compute an appropriate sample size to detect an actual effect with high probability. Another application would be to determine the alpha level (α) based on the other established parameters. This application is less frequent due to various reasons, some of which will be explained later. Finally, it can be used to calculate the effect size as a function of the other elements in the formula, that is, α, N, and power. This application is also relatively rare (Cascio & Zedeck, 1983).

Instead of giving a didactic example for each of these applications, the section entitled "Empirical example" describes a typical research-planning sequence. It covers various applications of statistical power and, therefore, is an exhaustive illustrative means to explain the various aspects. Before detailing the research-planning protocol, let us consider the general sequence and define the concepts used.

Concepts related to statistical power

Statistical power in the research-planning phase

In order to maximize the power of a test, Cascio and Zedeck (1983) recommend following this sequence in the research-planning phase (*a priori* application):

1. Determine the minimum effect size that would be considered useful or significant.
2. Determine the appropriate sample size based on the desired power, the selected effect size and the given alpha level.
3. For a fixed sample size that proves insufficient to achieve the desired power given the other parameters, adjust the alpha level while considering the relative impact of type I and type II errors.

But first, let us review some of the concepts involved.

Significance levels α (alpha) and β (beta)

In the part entitled "Statistical Power, definition and application", we said that, for a certain effect size (e.g. d) and a given sample size (N), the significance level α allows to quantify the power of a test, and vice versa. For the purpose of such a statement, in order to ensure that "vice versa" fully applies, it is imperative to consider the α level as being variable. Most researchers believe that this significance level must be set at $\alpha = .05$; but why not at .08, .10 or .025?

The conventional $\alpha = .05$ is widely believed to have been established, more or less arbitrarily, by Sir Ronald Fisher (Sedlmeier & Gigerenzer, 1989; Ryan, 1985; Cohen, 1990). Indeed, Fisher considered that variations relative to normal that are greater than two standard deviations (which roughly corresponds to $\alpha = .05$ for two-tailed tests) must be judged significant (Fisher, 1925). Later, Fisher stated that he personally *preferred* to set a low significance level and reject the results that did not meet this criterion (Fisher, 1926). In this context, the word *prefer* is of great importance. In fact, this preference has been questioned and challenged by many accomplished statisticians who described the unconditional use of $\alpha = .05$ as an "almost religious extreme" (Cascio and Zedeck, 1983), as "sacred" (Skipper, Guenther, & Nass, 1967), as an "arbitrary unreasonable tyranny" (Cohen, 1990), or as "decreed by tradition and reviewers" (Tabachnick & Fidell, 2001).

This debate is even more pertinent nowadays since today's statistics theory differs from Fisher's teaching. As stated by Sedlmeier and Gigerenzer (1989), the current theory is a hybrid of approaches developed by Fisher, on the one hand, and by Neyman and Pearson, on the other hand (see below). Therefore, Fisher's preference for α = .05 has been transposed to a different context and, while well intentioned at the beginning, it no longer corresponds to the current reality. The following paragraphs explain the controversy surrounding the Fisher and the Neyman – Pearson approaches.

At the time of Fisher, and largely on his recommendation, solely the null hypothesis (H_0: the hypothesis that we formulate with the hope of rejecting) was specified and verified (Fisher, 1935, 1966). Such verification consists of computing the value of a statistic (for example, t or F) based on the results, while considering the null hypothesis as being true. The value of this statistic is then compared to the critical value, which is also computed using a certain probability level α. If the statistic value is higher than the critical value, it is presumed that the results concerned have not arisen through chance and H_0 is therefore rejected. If H_0 is rejected when in fact it is true, then a type I error has been committed (rejecting H_0 when H_0 is true).

However, working in the shadows of Fisher, the Neyman – Pearson team had already developed the concepts of alternative hypothesis (H_1) and of type II error (i.e. rejecting H_1 when H_1 is in fact true, or not rejecting H_0 when H_1 is true). In other words, the alternative hypothesis H_1 is a statement in the favor of which the null hypothesis could be rejected. These concepts allow to compute the probability of committing a type II error, denoted by β. In other words, it is the probability that the alternative hypothesis would be rejected in favor of the null hypothesis when H_1 is in fact true; that is, the probability that a true effect (H_1) would be considered the result of chance alone and hence judged false. The next step would be to compute $(1 - \beta)$, or the probability of an effect being found true when it is in fact true. Therefore, $(1 - \beta)$ represents the power of a test to detect a significant result when the effect actually exists.

While being of great value, the Neyman – Pearson theory did not achieve the desired impact, perhaps due to the strong opposition on the part of Fisher who described those interested in the concepts of type II error and statistical power as "Russians trained for technological efficiency rather than statistical inference" (Fisher, 1955). This battle of opinions, which passed relatively unnoticed in North America, had far-reaching consequences. The main outcome was a new hybrid theory (as described by Sedlmeier & Gigerenzer, 1989) taught in Human Sciences programs nowadays. This hybrid theory has been approved neither by Fisher followers nor by Neyman – Pearson supporters. The hybrid theory states that only the null hypothesis must be verified, as recommended by Fisher, but also recognizes the importance of the type II error, as suggested by Neyman and Pearson, therefore allowing to set different values for α before gathering test data (some reports use α = .10). However, this hybrid theory refers to the type II error and statistical power solely for general academic purposes since their calculation calls for an alternative hypothesis (H_1) that is not part of the theory.

The hybrid theory could be improved by considering the β level – and thus statistical power – in order to determine the α level (which, of course, implies that its value could vary). In fact, many statisticians (e.g. Cascio & Zedeck, 1983) recommend making a beforehand evaluation of the relative impact of the type I and type II errors on the desired power (note that this approach is not unanimously accepted, see Ryan, 1985). For example, a researcher wants to maintain a typical statistical power of .80, then by definition β = .20. If this researcher chooses the traditional α = .05, then $\beta / \alpha = .20 / .05 = 4$, which means that a type I error would be considered four times as significant and harmful as a type II error (reasonable). Let us consider another example. A researcher believes that a lower α would yield a better test, and chooses α = .001. According to Cohen (1988), such a weak α level is typically associated with a very low statistical power, for example, power = .10. Then, β = 1 – power = 1 – .10 = .90 and $\beta / \alpha = .90 / .001 = 900$. It means that, according to this researcher, a type I error would be 900 times as critical and harmful as a type II error. With certain exceptions, such relative importance of errors indicated by this researcher seems rather unreasonable…

Therefore, the values of α and β levels may have a significant impact on the results of the statistical tests used (type I and type II errors). Thus, their respective importance is determined by a "quantity" that, by definition, is closely related to the concerned data: effect size. We shall discuss it later.

Effect size of the observed phenomenon

In 1988, Cohen stated that effect size is the least known concept related to statistical inference. He attributed such relative obscurity to the historical difference between Fisher's testing philosophy and Neyman and Pearson's (1928, 1933). In fact, Fisher's test procedures do not include a defined alternative hypothesis, which makes it impossible to calculate the β probability and, consequently, the statistical power of a test. To do so, we need to formulate H_1, which implies a certain degree of the effect presence in the population and/or a certain degree of falsity of the null hypothesis. Then, such degree is indeed the effect size.

More specifically, when comparing two populations (i.e. inter-group tests), the null hypothesis usually takes the following form: "the difference between the measured parameters for each population is zero", or $\mu_2 - \mu_1 = 0$. Therefore, if the null hypothesis is true, the effect size has to be zero. Consequently, if the null hypothesis is false, then $\mu_2 - \mu_1 \neq 0$. That would be similar to recognizing the existence of a difference between the means of the two populations, i.e. $\mu_2 - \mu_1 = x$, where x is the effect size or, in other words, the "degree of falsity" of the null hypothesis (H_0). The higher the value of x, the farther the null hypothesis is from the truth. Note that the specification of x, as per Neyman and Pearson, is equivalent to specifying H_0: $\mu_2 - \mu_1 = 0$ and H_1: $\mu_2 - \mu_1 = x$.

We see that the equation $\mu_2 - \mu_1 = x$ defines x in terms of the unit scale used (e.g. seconds, IQ points, etc.). So, if we want to use power charts or compare the results of several tests, the effect size must be specified as a dimensionless number. Depending on the actual test, the effect size may be expressed as d (difference between two means), r (correlation between two variables), f (ANOVA test) or any other index related to the specific test (see Cohen, 1992). For the purpose of an example, the d formula is:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

If we divide the difference between the means measured on a given scale by the standard deviation expressed in the same units of measurement, we see that the effect size d is indeed independent of the scale used. The same is true of the other effect size indices (e.g. r, f, etc.).

Since d is defined as the difference between two means divided by a standard deviation, it can be easily computed once the sample data has been collected. However, as mentioned earlier, the most common application of the statistical power concept is in the research-planning phase, a phase when sample means are not yet available. In this case, we have to find other ways to make a realistic estimate of the effect size (including useful or significant effect, as specified in the sequence by Cascio & Zedeck, 1983). Cohen (1988) suggests two approaches: 1) one can calculate the effect size from previous work in a similar area (e.g. meta-analysis studies) or 2) if such data are not available, one can use personal judgment, theoretical principles or any combination thereof to estimate the possible effect size in the study. As an alternative to the first approach, many funding agencies now encourage researchers to realise pilot studies in order to test the feasibility of the research protocol and to obtain preliminary effect sizes using the specific measures intended for the main study. However, it remains a good idea to base preliminary effect size estimations on prior work. Papers by Levine (1997) and Thalheimer and Cook (2002) provide interesting introductions on the subject. If using the second approach, the researcher will identify the anticipated effect size as conventional values of "small", "medium" or "large". In Human Sciences, an effect size is defined as "medium" if it is perceptible to the naked eye of an attentive observer, as "small" if it is significantly smaller than the medium effect size, but not trivial, and as "large" if it is larger than the medium effect size, and the difference between the two is similar to that between the small effect and the medium effect (see Cohen, 1992). Mathematically speaking, effect sizes are defined as "small" if d = .20, "medium" if d = .50, and "large" if d = .80.

Before we go back to the research design sequence described by Cascio and Zedeck (1983), there is still one more point to address. We need to specify the relationship between power and sample size.

Sample size

The most frequent application of power analysis is to compute the minimum sample size (N) required to test an effect of the estimated size with a desired power and a known α level. Generally, a larger sample size tends to reduce the variability of sample statistics (mean, correlation, etc.), or in other words, reduces error variance and therefore increases the likelihood of detecting an effect size of the specified (or larger) magnitude. From the statistics point of view, it reduces the β probability and therefore increases statistical power ($1 - \beta$).

Unless impossible due to major constraints, sample size must be the first criteria to be adjusted in order to augment power. However, Tabachnick and Fidell (2001; p. 35) caution researchers against excessive use of a large sample size (N), which may cause the statistical power of a test to be too strong. In fact, in such cases, the null hypothesis would be almost certainly rejected and the test might be able to detect effects that are too small to be of any substantive significance. In a way, the fact that journals now insist on reporting effect size estimates along statistically significant results would tend to minimize the impact of such findings.

Empirical example

To illustrate the research design sequence by Cascio and Zedeck (1983) as well as its related concepts, here is an example that demonstrates recurring concerns in clinical research. The example is purely fictional.

Suppose a researcher wants to investigate the effectiveness of a short-term (14 weeks) psychodynamic therapy treatment of minor depressive disorder (introduced for further study in the DSM–IV and DSM-IV-TR; American Psychiatric Association, 1994, 2000). In her study, she chooses to use the Beck Depression Inventory – II (BDI-II, 1996) to compare the patients' scores at the end of the psychotherapy treatment (posttest) with their scores at the beginning of the psychotherapy treatment (pretest).

For this project, the researcher opts for the planning sequence suggested by Cascio and Zedeck (1983). The priority is thus to define the minimum effect size that would be judged useful or important. She reads on the subject and concludes that, based on the little data available, individuals classified as suffering from minor depression have an average initial BDI-II score of about 19 with a standard deviation of 8. She also learns that in a nonclinical population, this BDI-II score is usually around 8 with a standard deviation of 6 (these last figures, however, will not be part of the calculation). Finally, she concludes that the initial BDI-II score is not a strong predictor for the posttest score in the treatment of severe depression and that the correlation between the two scores is merely .20.

Suppose now that the researcher considers that a final mean BDI-II score of 15, compared to the initial score of 19, would indicate that the treatment is somewhat effective and that it is worth pursuing. On the other hand, an improvement of less than 4 points by the end of the treatment could lead her to abandon this treatment method and to reconsider its pertinence.

Since the pretest and posttest data are part of a repeated measures design and thus are not independent (for further details, see Howell, 1998, part 8.5), the researcher computes the effect size as follows:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{X_1 - X_2}} = \frac{\mu_1 - \mu_2}{\sigma\sqrt{2(1-\rho)}}$$

Since the posttest standard deviation is not available, she estimates that such deviation should roughly correspond to the pretest value. Hence:

$$d = \frac{19 - 15}{8\sqrt{2(1 - .20)}} = .40$$

Further, being familiar with power analysis, the researcher believes that it would be wise to maintain a minimum test power of .80. Moreover, she knows that using a standard α level of .05 in her tests would mean better chances to have her work published. In Howell's table (1998; p. 762) she finds a corresponding δ (the so-called *noncentrality parameter*) which in this case is 2.80. She substitutes these values in the following formula:

$$\delta = d\sqrt{N},$$

hence

$$N = \frac{\delta^2}{d^2} = \frac{2.80^2}{.40^2} = 49$$

The calculation shows that she needs to recruit 49 participants to have an 8 in 10 chance to obtain a significant result with $\alpha = .05$ if the treatment yields an effect size of at least .40.

Suppose now that she wishes to investigate the effect of her treatment on patients who show comorbid personality disorders (e.g., borderline, histrionic, and narcissistic) and that she has access to a maximum of 30 participants for her study. Due to its preliminary nature, there is no indication whether the considered treatment could be effective in the case of such patients. In addition, the presence of comorbid personality disorders is likely to make the group more heterogeneous than one may conclude based on initial BDI-II scores. Consequently, instead of α / β of 1 in 4, the researcher opts for an α / β ratio of 1 in 2. If she is determined to maintain a power of .80, then $\beta = .20$ and $\alpha = .10$, which corresponds to $\delta = 2.50$. If we redo the earlier calculation using these new criteria, we see that the number of participants in this case should be 39. The researcher is now 10 patients short.

Under the circumstances, the researcher has several options. First, she may decide to postpone her research. Second, she may try to team up with colleagues working in the same field in order to increase the number of participants available for inclusion in the study. Third, she may proceed with the study if she accepts statistical power below .80 (in the last test design described above with $N = 30$ and $\alpha = .10$, power will be around .70). In this case, the researcher could try to limit her sample to a particular subgroup in order to reduce within-group variance. Finally, she may opt for a one-tailed *t*-test. However, it should be noted that one-tailed tests are not unanimously accepted and should be used sparingly and with great caution.

Consequences of power analysis

Statistical significance vs. design quality in reviewers' decision

In 1982, Atkinson, Furlong and Wampold (1982) concluded that a study report submitted for publication was most likely to be rejected unless at least some of its major results were significant at traditionally accepted levels $p < .05$ or $p < .01$ (see also Sedlmeier & Gigerenzer, 1989). Perhaps due to the impact of this study, the subject of the statistical significance of findings has become somewhat of a censorship gage (personal censorship on the part of authors themselves or censorship on the part of publishers, who knows? For an interesting perspective on this question, see Reysen, 2006). It remains difficult to conclude if the situation has ever improved since 1982. If we review the bits and pieces of information from various published sources (e.g., DeVaney, 2001), it seems that reviewers' position has somewhat improved since the 1982 article by Atkinson et al., but unfortunately not that much.

Without focusing too much on this subject, we would like however to point out that the concept promoted by reviewers and editors, as described by Atkinson et al., is in contradiction with the third step of the research design sequence by Cascio and Zedeck (1983; see Part 2). In fact, reviewers' position, just like the old Fisherian approach of $\alpha = .05$, does not allow for α to be considered as a variable. We share the views expressed by Sedlmeier and Gigerenzer (1989) and believe that promoting the concept of statistical power could shake the foundations of such status quo. One of the benefits of the statistical power concept is that it offers a rigorous rational approach that justifies the use of a variable α. Another advantage is that the statistical power concept encourages researchers to define beforehand a minimum effect size that would be judged useful or significant. According to Cohen (1994), careful attention to the effect size will result in reconsidering error variance (a smaller error variance is desirable), which should in its turn improve experimental designs.

Overestimation of effect size in the literature

Reviewers' bias towards studies reporting significant results has yet another consequence: studies with larger effect sizes are much more likely to be accepted for publication since they report significant results more often. Therefore, Lane and Dunlap (1978) point out that in such studies completed in a low statistical power context, reported effect sizes may be much higher than they are in reality. They explain that low α levels (e.g. α = .01) used in reports may result in distorted and artificially inflated effect sizes. The authors conclude that the general trend to publish significant results only cannot coexist with the adequate estimate of effect sizes based on the literature (in particular, with regard to meta-analyses). Consequently, they recommend accepting for publication all experiments if they relate to important concepts and have a well-structured design.

While this issue has been addressed in many publications, the work by Lane and Dunlap (1978) stands out since it spotlights the paradox that exists to date, i.e. the gap between the use of meta-analysis results to compute power, on the one hand, and publication requirements, on the other hand. To resolve the issue, the literature must mirror the real world more accurately. And to achieve that, we must question the conventionally established α level and promote studies of higher statistical power that are designed around the concept of minimum effect size.

Conclusion

To sum up, it seems that statistical power and its derived concepts have gained ground compared to the situation that existed several decades ago. However, the new awareness exists mostly in theory since the power of recent studies does not seem to differ much from the statistical power of studies reported by Cohen in his 1962 article. Given the positive impact that the use of statistical power could have on the scientific literature in general and on research planning in particular, we hope that its popularity would go from theory to practice. More specifically, we should go back to the sources and reconceptualize the type II error based on the concept of minimum effect size (i.e. minimum effect judged useful or significant). Moreover, we should stop regarding the α level as a constant (i.e. α = .05), as some magic limit that defines the truth. In fact, the α level should be considered as a variable, the value of which is determined by a realistic α / β ratio. Therefore, the statistical power of tests should be maintained above minimum to ensure that nonsignificant results could be considered as meaningful. In the context of powerful tests, nonsignificant results are in fact very pertinent since they suggest that the obtained effect sizes are relatively trivial (very small) or negligible (below the minimum effect judged useful or significant). Finally, reviewers should readily accept to publish significant and nonsignificant results alike to make sure that the reality depicted in the literature actually corresponds to the reality we know.

References

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th Ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders – Text Revision* (4th Ed.). Washington, DC: American Psychiatric Association.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29,* 189-194.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd Ed.). San Antonio, TX: Psychological Corporation.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23*, 399–406.

Cascio, W. F., & Zedeck, S. (1983). Opening a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology*, *36*, 517-526.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology*, *88*, 71–83.

Cohen, J. (1962). The statistical power of abnormal – social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

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

Denton, F. T. (1990). The effects of publication selection on test probabilities and estimator distributions. *Risk Analysis, 10,* 131-136.

DeVaney, T.A. (2001). Statistical significance, effect size, and replication: What do the journals say? *Journal of Experimental Education, 69,*310-320.

Fagley, N. S. (1985). Applied statistical power analysis and the interpretation of nonsignificant results by research consumers. *Journal of Counseling Psychology, 32,* 391-396.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd..

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33,* 503-513.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B, 17,* 69-78.

Fisher, R. A. (1966). *The design of experiments* (8th Ed.). Edinburgh, Scotland: Oliver & Boyd (1st Edition published in 1935).

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance.* Cambridge: Cambridge University Press.

Howell, David C. (1998). *Méthodes statistiques en sciences humaines*. Paris: DeBoeck Université.

Instructions to Authors, Journal of Consulting and Clinical Psychology (2007). Retrieved September 23, 2007, from http://www.apa.org/journals/ccp/submission.html

Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36*, 212–219.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107-112.

Levine, J. (1997). Overcoming feelings of powerlessness in "Aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12,* 84-106.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology, 20*, 76–78.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49*, 103–120.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrica, 20a,* 175-240, 263-294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Transactions of the Royal Society of London Series A, 231,* 289-337.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making type I errors. *Psychological Bulletin, 102,* 159-163.

Reysen, S. (2006). Publication of nonsignificant results: A survey of psychologists'opinions. *Psychological reports, 98*, 169-175.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646-656.

Ryan, T. A. (1985). Ensemble-adjusted *p* values: How are they to be weighted? *Psychological Bulletin, 97,* 521-526.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 2,* 16-18.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics* (4th Ed.). Boston, MA: Allyn and Bacon.

Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research articles: A simplified methodology*. Retrieved January 30th, 2007 from http://work-learning.com/effect_sizes.htm.

Manuscript received September 26th, 2006