Evaluating and Reporting Statistical Power in Counseling Research

Richard S. Balkin and Carl J. Sheperis

Despite recommendations from the *Publication Manual of the American Psychological Association* (6th ed.) to include information on statistical power when publishing quantitative results, authors seldom include analysis or discussion of statistical power. The rationale for discussing statistical power is addressed, approaches to using G*Power to report statistical power are presented, and examples for reporting statistical power are provided.

Despite existing differences between individuals who research counseling phenomena and those who practice in the profession, the American Counseling Association (ACA, 2005) adopted an ethical code that promotes the use of evidence-based interventions by persons who practice counseling. Thus, counselors must now be competent consumers of research conducted in the counseling profession and be able to make decisions about the applicability of research results. One of the foundational principles that counselors should understand when evaluating research is statistical power. Thompson (2006) classically defined statistical power as "the probability of rejecting the null hypothesis when the null hypothesis is false" (p. 172). In other words, power is simply the likelihood of finding a statistically significant result, if such a result actually exists, and is considered an essential element in designing and evaluating quantitative findings for research in counseling (Granelllo, 2007). Our intention is to provide readers with an understanding of the rationale for including power analysis in counseling research, the principles of power analysis, the various types of power analysis, and the methods for calculating power using computer software. We believe that by developing an understanding of power, readers will become better consumers of research. We also believe that the development of guidelines for power analysis will lead to more consistent reporting of power by counseling researchers.

**Rationale for Power Analysis**

Counseling researchers should consider the desired effect size, significance level, and appropriate sample size before conducting quantitative studies (Onwuebuzie, 2004). However, power is rarely reported in social science research (Cohen, 1992; Onwuebuzie, 2004), and the consequences for journals not employing this practice can be serious. Low statistical power contributes to an increased likelihood of making a Type II error (Onwuebuzie, 2004), thereby causing important findings to be either misreported or not even published. Although statistical procedures are common to protect the researcher from making a Type I error (finding a statistical relationship when one does not exist), such as by using a conservative alpha level as .01 or .001, less attention is often paid to decreasing the chance of making a Type II error (failing to find a relationship when one actually does exist).

To bridge the gap between research and practice, we illustrate the implications of misinterpreting research findings for the practicing counselor. For example, a school counselor is asked to develop a peer-helper program for the school district. In line with the *ACA Code of Ethics* (ACA, 2005), the school counselor reviews the research related to peer programs. Attempting to follow guidelines for best practice, the school counselor rejects a potential peer program because the researchers underestimated the magnitude of their results and failed to report power in their article, analogous to a Type II error. Because the researchers misrepresented the data, the school counselor misses the opportunity to take advantage of a potential resource that could make a difference in the lives of the children in the district, and the students are denied an opportunity to receive an effective program. Therefore, the necessity of implementing procedures to protect against Type II error may be equally important to procedures to protect against Type I error. Furthermore, we believe that both counseling researchers and counseling practitioners should understand these procedures to ensure a check-and-balance system. Thus, our primary focus for this article is to illustrate the concept of statistical power for all counselors and counseling researchers.

In the past, researchers have relied on imprecise guidelines to address sample size issues, rather than engaging in power analysis to determine appropriate sample size before data are collected. Researchers may find power analysis too complicated and may simply avoid the issue (Cohen, 1992). However, the advent of free software to conduct power analyses has simplified the process quite extensively. Specifically, G*Power software, which is compatible with both Mac and PC systems, is a free software program to assist researchers in conducting power analyses either a priori or post hoc (Mayr, Erdfelder, Buchner, & Faul, 2007). Following the heuristic examples are overviews of how G*Power can be used to...
conduct power analyses. More in-depth tutorials are available from the G*Power website.

Principles of Power Analysis

One reason that tests of statistical significance should not stand alone is because of the influence of sample size. With differences in magnitude between groups and level of significance held constant, changes in sample size will affect statistical power. To understand the influence of sample size, we use a conceptual illustration of basic statistical formulae. We recognize that many readers will not have committed these principles to memory. Thus, as a refresher, recall that parametric statistics, that is, statistics based on a probability distribution (e.g., a normal curve), follow the same basic formula: mean differences divided by error. Basic examples of this pattern may be found in the z-test, t-test, or analysis of variance (ANOVA) formulae and may be expressed as a fraction in which the numerator represents mean differences and the denominator represents error (Balkin & Erford, 2008). The error term (the denominator) incorporates two elements: (a) some type of variability (e.g., standard deviation, variance, sum of the squared deviations, or other error term) and (b) sample size. If the sample size increases, the value of the error term (the denominator) decreases. Assuming that the mean differences (numerator) are held constant, the ratio of the mean differences to the error increases, and the likelihood of finding statistical significance increases. The corollary is also true. If the sample size decreases, the value of the error term increases. Assuming that the mean differences are held constant, the ratio between the numerator and the denominator decreases, and the likelihood of finding statistical significance decreases.

In summary, larger sample sizes lead to less error and higher statistical power. Intuitively, larger sample sizes should lead to better estimates of the target population parameters (e.g., means, correlation coefficients) and, therefore, a decrease in sampling error.

This mathematical concept illustrates that the likelihood of finding statistical significance (i.e., rejecting the null hypothesis) increases with larger sample sizes, and the likelihood of failing to find statistical significance increases with smaller sample sizes (i.e., retaining the null hypothesis). Power, therefore, is related to Type II error. When sample size is small, the likelihood of finding statistical significance decreases, and researchers are in danger of making a Type II error—failing to find statistical significance when it actually exists. Because small sample size may lead to increased error, statistically significant relationships may not be evident, even if the mean differences are quite large. When researchers encounter a statistically nonsignificant result with moderate to large effect sizes, a Type II error is usually evident. Simply put, when nonsignificant finding are evident, statistical power should be evaluated to ensure that a sufficient sample size was used for the study.

Similar to the effect that the error term has on statistical significance, the magnitude of the mean differences or relationships (effect size) also influences statistical significance. The relationship between power and effect size can be understood both theoretically and mathematically. Thus, statistical power increases when the effect size (the numerator) increases and the error term is held constant. An elaboration on the role of effect size is outside the scope of this article; interested readers may refer to Trusty, Thompson, and Petrocelli (2004) for further study.

Another way that researchers can affect power is to adjust the level of significance. When more conservative alpha levels are used in statistical analysis (e.g., .01 or .001), statistical power decreases. Researchers are much less likely to find statistical significance and more likely to make a Type II error but gain the benefit of avoiding a Type I error. So why not use more liberal alpha levels and settle this issue? Remember that the level of significance, alpha, is a measure of the amount of Type I error in a study, and there is a relationship between power and error. Thus, researchers must be cautious about using a too-liberal alpha level. As consumers of research, readers should recognize that establishing an alpha level is a subjective decision by a researcher. As alpha increases, power decreases, but so does the likelihood of making a Type I error; conversely, the likelihood of making a Type II error decreases. The corollary is also true. As alpha decreases, power decreases, the likelihood of making a Type I error decreases, but the likelihood of making a Type II error increases.

As consumers of research, investigators must use careful thought when determining an alpha level, and some factors to consider include the level of significance in previous research and the ability to obtain a sufficient sample size. For example, researchers often seek to replicate the results of someone else's work. If results can be replicated, then there can be more confidence in the findings. As a rule of thumb, researchers seeking to replicate a study may wish to use a more conservative level of significance. By doing this, the researchers are setting a higher standard for their research outcomes. As we indicated earlier, the researchers in this particular case have the potential to decrease power by using a more stringent alpha level. However, such complications can be avoided when a sufficient sample size is used. Researchers have limited control over effect size and level of significance, but identifying and collecting an adequate sample size may be within their control.

Types of Power Analysis and Software Applications

As is the case with all types of statistical analyses, there is more than one method determining power. To help readers understand power in reported studies, we discuss the principles of and rationale for a priori power analysis, post hoc power
analytic, and sensitivity power analysis. We also include methods for using *G*Power under each type.

For readers who are engaged in counseling research, we offer some direction in the use of *G*Power as a software application for conducting power analysis. *G*Power is a free power analysis program for PC and Mac computers from the University of Trier, Germany. The *G*Power website includes information for free downloading, manuals, literature, and tutorials (http://wwwpsycho.uni-duesseldorf.de/aap/projects/gpower/). The program provides power estimates related to the type of statistical analysis. Users enlist drop-down menus to select the type of power analysis, statistic, and relevant components to perform power estimates. A description of using *G*Power follows.

**A Priori Power Analysis Using *G*Power**

When counseling researchers want to determine appropriate sample sizes, an a priori power analysis is indicated. If the researcher can estimate the magnitude of the desired effect to be detected (e.g., the desired degree of difference between the sample and the population) and the level of significance in the study (e.g., .05), the sample size necessary to find statistical significance can be computed using *G*Power. Cohen (1988) suggested that the Type II error risk should be 4 times as great as the Type I error risk to ensure adequate analyses without having to use unrealistically high sample sizes for social science research. Assuming an alpha level of .05, the recommended adequate power is .80, indicating that levels lower than .80 increase the chance of a Type II error to greater than 20% and higher levels of power may mandate the use of unrealistically high sample sizes.

To conduct an a priori analysis using *G*Power, researchers should consider (a) the desired effect size to determine statistical significance, (b) the alpha level of the study, (c) the desired amount of power, and (d) the number of groups or predictors in the study. For more complex designs (e.g., repeated measures and multivariate procedures), additional information such as the correlation among repeated measures and the nonsphericity correction may be necessary. *G*Power users can see the requirements necessary for each test by clicking on the drop-down menu labeled *Test family* and selecting the appropriate statistical test from the drop-down menu labeled *Statistical test*. Under *Type of power analysis*, select *A priori: Compute required sample size given α, power, and effect size*. In determining the desired effect size, *G*Power allows the user to place in an exact effect size or use Cohen's (1992) classification of small, medium, or large effect size as thresholds. Note that to detect more minute differences or relationships (i.e., small effect size), a larger sample size will be necessary. The alpha level selected is simply based on the preference of the researcher(s), taking into account the consequences of using a more conservative or liberal alpha level. Recall that larger sample sizes are necessary when more conservative alpha levels are used. In determining the desired amount of power, *G*Power uses a default value of .95, but users can change it manually to any other value, such as .80 to conform to Cohen's guidelines. The additional criteria are entered manually depending on the design of the study. Values for correlations and sphericity may be estimated based on previous research or researcher knowledge of the constructs. By selecting *Calculate*, the total sample size necessary for adequate power as specified using the above parameters is calculated and displayed.

Researchers who conduct an a priori analysis can make statements in their Method section related to the rationale of the sample size for a particular study. Some authors offered guidelines to determine sample size independent of a priori power analysis (e.g., Tabachnick & Fidell, 2001), but Pedhazur (1997) cautioned against such guidelines:

Sound principles of research design dictate that the researcher first decide the effect size, or relation, deemed substantively meaningful in a given study. This is followed by decisions regarding level of significance (Type I error) and the power of the statistical test (1 - Type II error) . . . meaningfulness cannot be arrived at in a research vacuum. (p. 26)

As an example of the utility of a priori power analysis, consider a researcher using *G*Power who wishes to conduct a multiple regression analysis with three predictor variables. In this case, the researcher identifies that a medium effect size ($R^2 = .13$; Cohen, 1992), with an alpha level of .05 and power at .80, would require a sample size of at least 77 participants. In the Method section of the study, the researcher could state, “An a priori power analysis was conducted using *G*Power. With an alpha level of .05, minimum power established at .80, and a moderate effect size of .13 (Cohen, 1992), 77 participants would be necessary to find a statistically significant effect in the model.” Thus, the researcher has been able to provide justification for the sample size in the study.

**Post Hoc Power Analysis Using *G*Power**

We advocate for a priori power analysis in counseling research, but there are times when there is a lack of extant research regarding a specific research topic and estimating effect size may not be possible. However, failure to conduct a priori power analysis may not always have a clear rationale. We concur with Onwuegbuzie (2004), who indicated that the lack of a priori power analyses may be due to a lack of understanding by researchers about the relevance of power analyses, inconsistent American Psychological Association standards for conducting and reporting statistical power, and the inability of researchers to conduct power analyses in commonly used statistical programs, such as SPSS and SAS. Researchers should be cautious about using post hoc power analysis procedures. Authors are encouraged to consider levels of effect size to detect statistical significance and the appropriate sample size for identifying such a magnitude to avoid findings that may be
less meaningful. Researchers thoroughly review the literature to make educated estimates for effect size in a priori power analysis. When such estimates are not warranted because of limited evidence of effect sizes in previous research, researchers should use conservative estimates (Cohen, 1988, 1992).

Unlike an a priori power analysis that relies on estimates of effect size, post hoc power analyses convey the actual power in the study through the observed effect size rather than an estimated value. The advantage of this method is that the occurrence of a statistically nonsignificant finding may be evaluated. Statistical nonsignificance may occur because of insufficient power resulting from (a) an insufficient sample size and/or (b) a small effect size. A study with a large effect size showing statistical nonsignificance would be an inconclusive finding, because meaningful differences should occur with statistical significance. A statistically nonsignificant finding with a small effect size and adequate sample size is a conclusive finding (Onwuegbuzie, 2004), because nonmeaningful differences are evident. Such a finding could contribute to existing research, as the importance of knowing variables and interventions that may contribute to change is as important as knowing what interventions and variables do not relate to a specific phenomenon.

The procedure for conducting a post hoc power analysis using G*Power is the same as for the a priori analysis with the following exception. Under Type of power analysis, select Post hoc: Compute achieved power given α, sample size, and effect size. The researcher can now provide the exact effect size obtained in the study. By selecting Calculate, the achieved power of the study is displayed. Additionally, achieved power can be evaluated in current versions of SPSS.

As we indicated at the beginning of this article, journals have been inconsistent in requiring researchers to report power. As such, consumers of research may make determinations about the utility of results without all of the relevant information. Based on our description of post hoc power analysis using G*Power, it is possible for readers to make their own determinations of power (given adequate descriptive statistics) even when researchers fail to provide it. For example, Chen, Chou, and Yang (2005) found a statistically significant difference in health responsibility among underweight, average, and overweight Taiwanese adolescents (N = 583). However, no effect size or power analysis was given in the article. Based on the descriptive statistics provided in the study, we calculated Cohen’s f² to be .14. Using G*Power and identifying an alpha level of .05, an effect size of .14, and a sample size of 583, we found the post hoc statistical power to be .86. Given this new information from the post hoc power analysis, findings reported by Chen et al. (2005) were not very meaningful; the large sample size was essential to the significant finding, as statistical power was more than adequate, but the effect size was small. Consumers of this research should use caution in interpreting the meaningfulness of their results. In a best-case scenario, Chen et al. would have included in their Results section a comment such as the following: "A post hoc power analysis was conducted utilizing G*Power. With an alpha level of .05, a sample size of 583, and a small effect size of .14 (Cohen, 1992), achieved power for the study was .86." Such a comment would identify the small effect size to the informed reader, thereby leading the reader to be cautious in interpreting the meaningfulness of the results.

Sensitivity Power Analysis Using G*Power

Another type of analysis that may be conducted after data have been analyzed is known as a sensitivity power analysis. The sensitivity power analysis provides the minimum effect size required to find statistical significance. Using G*Power, the sensitivity analysis requires an alpha level, sample size, number of groups, and the amount of power desired by the researcher in the study. So, in the Chen et al. (2005) study, a sensitivity analysis using .80 as a threshold for power, three groups, a sample size of 583, and an alpha level of .05 would yield an effect size of .129 to achieve statistical significance. A statement addressing the following may be used in the study: "Given a sample size of 583, an alpha level of .05, and minimum power of .80 (Cohen, 1992), there is an 80% chance of detecting an effect size of f = .129, assuming statistical significance and such an effect size actually exists.” The reader, therefore, obtains valuable information related to the magnitude of the difference or relationship required to obtain statistical significance with an 80% likelihood. Given the large sample in the Chen et al. study, only small differences were necessary to find statistical significance. In other words, the use of large samples is not always prudent and may actually skew the results of research. Without power analysis, it is difficult to determine the degree to which the results are affected by the sample size. Additionally, when researchers are committed to a specific sample size, due to cost constraints, time, funds, design, and so forth, a sensitivity analysis can be used to identify the type of effect size that may be detected with a reasonably high likelihood. Consumers and producers of research should use caution when interpreting statistical significance, without accounting for effect size and power (Onwuegbuzie, 2004; Trusty et al., 2004).

Summary

The American Psychological Association (APA) encourages power analysis for manuscripts to be published in APA journals. Researchers should “routinely provide evidence that the study has sufficient power to detect effects of substantive interest” (APA, 2010, p. 30). However, Onwuegbuzie (2004) noted that, to date, no examples or methods have been provided by APA to properly address this issue. Despite the lack of proper methodology for reporting power, there has been an increased emphasis of reporting power in counseling research. G*Power is a free and user-friendly software system that researchers may use to
address the standard of analyzing and reporting statistical power. G*Power provides technology to conduct a priori, post hoc, and sensitivity power analyses to ascertain the sufficiency of sample size for a study or whether adequate power is evident. Most importantly, G*Power can be used in research design to ensure that an adequate sample size is identified prior to data collection and analysis. Researchers waste valuable time attempting to conduct studies with inadequate power as a result of low sample size.

In this article, our purposes were twofold: (a) to inform the consumer of counseling research about the foundational principles of, rationale for, and interpretation of power analysis; and (b) to provide the producer of counseling research with methods and procedures for conducting and reporting various types of power analyses. We believe that the focus on power analysis will facilitate greater understanding of quantitative studies in counseling research.

References


