# Statistical Power Analysis

CHRISTOPHER L. ABERSON

## Abstract

Statistical power refers to the probability of rejecting a false null hypothesis (i.e., finding what the researcher wants to find). Power analysis allows researchers to determine adequate sample size for designing studies with an optimal probability for rejecting false null hypotheses. When conducted correctly, power analysis helps researchers make informed decisions about sample size selection. Statistical power analysis most commonly involves specifying statistic test criteria (type I error rate), desired level of power, and the effect size expected in the population. This article outlines the basic concepts relevant to statistical power, factors that influence power, how to establish the different parameters for power analysis, and determination and interpretation of the effect size estimates for power. I also address innovative work such as the continued development of software resources for power analysis and protocols for designing for precision of confidence intervals (aka, accuracy in parameter estimation). Finally, I outline understudied areas such as power analysis for designs with multiple predictors, reporting and interpreting power analyses in published work, designing for meaningfully sized effects, and power to detect multiple effects in the same study.

## INTRODUCTION

Many research findings involve application of inferential statistics. Inferential statistics refer to approaches used to draw conclusions about a population based on a sample. In short, we infer what the population might reasonably look like based on a sample. At the core of these approaches is the null hypothesis. The null hypothesis represents the assumption that there is no effect in the population. Common forms of the null might be that there is no correlation between two variables or that the means for two groups are equal.

The primary decision made through use of inferential statistics is a determination of whether the null hypothesis is a reasonable estimate of what the population looks like. More generally, we ask given what the sample looks like, could the population reasonably reflect the relationship stated in the null hypothesis? In making such a determination, errors are possible because

*Emerging Trends in the Social and Behavioral Sciences*. Edited by Robert Scott and Stephen Kosslyn.
© 2015 John Wiley & Sons, Inc. ISBN 978-1-118-90077-2.

samples may or may not represent populations accurately. A common analogy is to a jury trial. The trial begins with the presumption of innocence and a conviction requires proof beyond a reasonable doubt. The presumption of innocence is conceptually similar to the null hypothesis. A guilty decision requiring proof beyond a reasonable doubt means that given the evidence, it seems unlikely that the defendant is innocent.

As in a jury trial, hypothesis testing conclusions may be in error. Two possible errors exist. Falsely rejecting a true null hypothesis (the equivalent of a false conviction) and failing to reject a false null hypothesis (the equivalent of failing to convict a guilty defendant). Continuing the jury analogy, researchers usually take the role of the prosecution. They believe the defendant is guilty (i.e., the null hypothesis is false) and seek a conviction (i.e., rejection of the null hypothesis). Statistical power reflects a study's ability to reject the null. Power analysis maximizes the researcher's ability to attain that goal.

When researchers conduct power analysis, the purpose is usually to determine the sample size required to achieve a specified level of power. A research study designed for high power is more likely to find a statistically significant result (i.e., reject the null hypothesis) if the null hypothesis is in fact false than in a study with lower power.

## Null Hypothesis Significance Testing Terminology and Review

To understand power, a review of null hypothesis significance testing (NHST) is useful. NHST focuses on the probability of a sample result given a specific assumption about the population. In NHST, the core assumption about the population is the null hypothesis (e.g., population correlation is 0) and the observed result is what the sample produces (e.g., sample correlation of 0.40). Common statistical tests such as Chi-square, the t-test, and ANOVA determine how likely the sample result (or a more extreme sample result) would be if the null hypothesis were true. This probability is compared to a set criterion commonly called the *alpha* or *type I error level*. The type I error rate is the probability of a false rejection of a null hypothesis the researcher is willing to accept. A common criteria in the behavioral sciences is a type I error rate of 0.05. Under this criterion, a result that would occur less than 5% of the time when the null is true would lead to rejection of the null.

Table 1 summarizes decisions about null hypotheses and compares them to what is true for the data ("Reality"). Two errors exist. A type I or $\alpha$ error reflects rejecting a true null hypothesis. Researchers control this probability by setting a value for it (e.g., use a type I rate of 0.05). Type II or $\beta$ errors reflect failure to reject a false null hypothesis. Power reflects the probability of rejecting a false null hypothesis (one minus the type II error rate). Type II errors

**Table 1**
Null Hypothesis Testing Decisions and Errors

| **Reality** | | | |
|---|---|---|---|
| | | **Null hypothesis is true** | **Null hypothesis is false** |
| Statistical decision | Fail to reject null | Correct decision ($1-\alpha$) | Incorrect decision (type II or $\beta$ error) |
| | Reject null | Incorrect decision (type I or $\alpha$ error) | Correct decision ($1-\beta$ or power) |

are far more difficult to control than type I errors, as they are the product of several influences (see section on What Impacts Power). For readers wanting a more complete overview of NHST, most introductory statistics texts (e.g., Howell, 2011) provide considerable discussion of these topics.

### A Brief History of Statistical Power

Statistical power analysis came to prominence with Jacob Cohen's seminal work on the topic (e.g., Cohen, 1988). Since that time, an extensive literature and several commercial software and freeware packages focused on power and sample size determination (e.g., PASS, nQuery, Sample Power, G*Power) emerged. In recognition of the important role of power, grant applications often require or encourage statistical power analysis as do influential style manuals (e.g., American Psychological Association). Despite these advances and encouragements, surveys across numerous fields suggest that low power is common in published work (e.g., Maddock & Rossi, 2001). In the present chapter, I review some of the basic issues in power analysis, address factors that promote underpowered research, provide suggestions for more effective power analyses, examine recent advances in power analysis, and directions for the future.

## FOUNDATIONAL RESEARCH

### What Impacts Power (and Type II Error)?

The primary influences on power are effect size, type I error ($\alpha$), and sample size. Effect size reflects the expected size of a relationship in the population. There are numerous ways to estimate effect size; however, the most common measures are the correlation coefficient and Cohen's $d$ (difference between two means divided by standard deviation). In general, the correlation is most common when expressing relationships between two continuously scaled

variables and the $d$ statistic for discussing differences between two groups. Other estimates include $\eta^2$ and $\omega^2$ (both common for ANOVA designs), $R^2$ (common for regression), and Cramer's $V$ or $\varphi$ (common for Chi-squared designs). Regardless of the estimate of effect size, larger effect sizes produce greater power. Larger effects correspond to situations where the value for the statistic of interest is extreme compared to the value specified in the null hypothesis (i.e., no effect).

Type I error reflects a rate the researcher is willing to accept for falsely rejecting true null hypotheses (often 0.05 or 0.01). Accepting a higher type I error rate (e.g., 0.05 instead of 0.01) increases power as it makes it easier to reject the null hypothesis. Although increasing type I rates improve power, by convention researchers rarely go above 0.05.

Sample size influences power in a simple manner. Larger samples provide greater power. If the null hypothesis is, in fact, false, a larger sample size is more likely to allow for rejection.

Power analysis most typically involves specifying type I error, effect size, and desired power to find a required sample size. Specification of desired power (often 0.80 or 0.90) and type I error rate is straightforward. Effect size determination is not as easy. The effect size is "generally unknown and difficult to guess" and requires consideration of a wide range of factors such as strength of manipulation of variables, variability in dependent measures, and a meaningful magnitude of relationships (Lipsey, 1990, p. 47). Although complicated, effect size determination is likely the most important decision made in power analysis.

## Effect Size Determination

When we discuss effect size for power analysis, we are estimating what the population actually looks like. Of course, there is no way to know what the population looks like for sure. There are common standards for power (e.g., 0.80 or 0.90) and type I error rates (0.05 or 0.01), but there is no easy way to figure out what sort of effect size to expect for your study. Effect size estimation for power analysis requires careful consideration as this value influences the outcome of the power analysis more than any other decision. The more thought put into this estimate, the better the analysis.

*Small, Medium, or Large Effects.* Jacob Cohen's seminal work introduced concepts of effect size and power to several generations of researchers. In this work, Cohen provided numerous tables organized around finding sample size given desired power, effect size, and type I error. Later work simplified these tables to focus on small, medium, and large effects (and the sample sizes

needed to detect each) rather than exact effect size values. This work greatly advanced statistical power analysis and greatly increased the researcher's ability to address power. However, an unintended consequence of the work is that it appears to foster reliance on use of small, medium, and large effect size distinctions. Reflecting this, most empirical articles that report power analyses include a statement such as "a sample of 64 participants yields 80% power to detect a medium effect."

I believe that the small, medium, large distinction should be discarded. First, effect size measures combine two important pieces of information, size of difference between groups (or strength of association), and the precision of the estimate. Lenth (2000) provides a useful example that shows why considering differences and precision separately is valuable. In the example, two studies produce the same effect size (a medium effect in this case). In the first, a test detects a difference of about 1 mm between groups using an imprecise instrument ($s = 1.9$ mm). The second case involves a more precise measurement ($s = 0.7$ mm) and a within-subjects design. The second test allows for detection of means differences of around 0.20 mm. Both tests involve the same effect size but the second test is much more sensitive to the size of the differences. The focus on effect size may obscure important relationships.

Another pervasive issue with the use of small, medium, and large effect size distinctions is that their selection often fails to correspond to careful thought about the research problem. When I consult with researchers on power analysis, most tell me they designed for a medium effect (or a large effect) but few can tell me why they chose that effect size. It appears that these "shirt size" distinctions foster reliance on arbitrarily selected effect sizes. When effect sizes used in power analyses are arbitrary, the corresponding sample size estimates are meaningless.

## Choosing How Much Power

The choice of how much power is adequate for the research design usually reflects a combination of research and practical concerns. The primary research concern is the cost of making a type II error. For example, if a treatment were very expensive to develop, the cost of failing to reject the null hypothesis (i.e., having no evidence for effectiveness) would be high. In a case such as this, designing for considerable power (e.g., 0.95) would minimize the chance of such an error. This would likely come at the cost of a very large sample. In contrast, an experiment examining a more trivial relationship might reasonably settle on a more moderate level of power (e.g., 0.80).

In psychological research, there appears to be an unofficial standard of 0.80 for power. At first blush, this might seem low as it means accepting

a 20% chance of failing to reject false null hypotheses. The 0.80 criteria, however, reflects practical concerns over the optimal balance of sample size requirements and power. Generally, increasing power reflects consistent increase of about one-quarter of the sample size for moving from 0.5 to 0.6, 0.6 to 0.7, and 0.7 to 0.8. However, moving from 0.8 to 0.9 requires an increase of around one-third of the sample size. Getting from 0.9 to 0.95 requires another one-quarter increase. For example, if a sample size of 100 produced power = 0.50, then it would take roughly 25% more cases (25) to produce power = 0.60 (a total of 125 cases). Moving from 0.60 to 0.70 would require another 25% jump (an additional 31 cases for a total of 156) and moving from 0.70 to 0.80 reflects addition of 39 more cases (total cases 195). To get to power = 0.90, sample size must increase by about 64 (total cases = 258; a 33% increase in cases rather than 25%). This suggests that power of 0.80 provides the best balance between sample size requirements and power.

### When to Conduct Power Analysis

Power analyses should be conducted before data collection. Power analysis is an *a priori* venture that allows researchers to make informed decisions about sample size before beginning their work. In some cases, researchers do not have control over sample size (e.g., archival work). In those situations, it is reasonable to conduct power analyses that indicate the power to detect effects of various size (e.g., "the sample allows us 80% power to detect effects as small as $d = 0.25$) but this should not be presented as a justification for the sample size, only as information about limitations of conclusions drawn from the data.

Researchers should not conduct power analyses after completion of data collection. Some statistical packages (e.g., SPSS) provide computation of "observed power" based on samples. A common misuse of such values is found in statements such as "we failed to reject the null hypothesis; however, power for detecting effect was low, suggesting that a larger sample would allow us to support predictions." This statement is flawed. Any test that does not reach statistical significance (when the null is false) is underpowered. In fact, the probability produced by significance tests relates inversely to power. If your significance test probability is high, power will be low.

## CUTTING-EDGE WORK

### Software Advances

Early work provided numerous tables for determining statistical power, but power analyses at present are primarily software based. There are several commercial programs such as PASS (Power Analysis and Sample Size),

nQuery, and Sample Power. In addition, two freeware packages, G*Power and PiFace, provide an excellent array of analyses. R packages such as Pwr provide a range of analyses for simple designs and there are several packages addressing power for complex approaches (e.g., power Mediation, long power). In addition, the free Optimal Design package addresses power and sample size selection for multilevel models (aka hierarchical linear models). Although these resources provide accurate power analyses for many designs, it is important to recognize that solutions provide by software are only as good as the information provided by the researcher. For example, if the researcher provides arbitrary effect size values, the resultant sample size estimates will reflect that arbitrary decision.

### Precision Analysis/Accuracy in Parameter Estimation

Precision analysis, also known as accuracy in parameter estimation, determines sample sizes necessary to produce a confidence interval of a specified width. A confidence interval is an estimate of what the population reasonably might look like given our sample results. These intervals may be very wide (i.e., imprecise) or narrow (i.e., precise).

This approach fits with recent calls to focus on confidence intervals either in conjunction with traditional significance tests or in place of such tests. Recently developed approaches determine sample size requirements based on the desired precision of results (i.e., confidence interval width). The MBESS package for R provides an impressive array of protocols for precision for most statistical values (Kelley, 2007).

## DIRECTIONS FOR THE FUTURE

### Power Analysis for Complex Research Designs

Despite software advances, conducting power analysis for common, but complex research designs in the behavioral sciences is not well explicated. Take, for example, the case of a power analysis in multiple regression. The most common hypothesis tests for multiple regression focus on the squared multiple correlation ($R^2$; either for a model or for the addition of variables) and regression coefficients (aka, slope, b, beta). $R^2$ refers to the variance explained by all of the predictors in the model (or a specific set of predictors). The null hypotheses is that $R^2$ is 0.00 in the population (i.e., the predictors do not relate to the criterion variable). Regression coefficients reflect the variance uniquely explained by a predictor. That is, what a specific predictor explains that the others cannot. Each predictor's coefficient has a null hypothesis attached to it. The null in this case is that the individual predictor does not uniquely explain the dependent variable. I use multiple

regression as an example, but the general issues discussed are applicable to any analyses that employ multiple predictor variables.

Power analyses for $R^2$, in terms of models and change, are handled well by many applications (e.g., powerreg command in STATA, proc power in SAS, G*Power). $R^2$ values are straightforward to address, requiring only information about the proportion of explained variance and the number of predictors.

The calculation of power for a coefficient is more complex. Power for coefficients is a function of both their relationship to the criterion measure and their relationships with each other (i.e., correlation with other predictors in the model). Power decreases as a predictor's overlap with the other predictors increases. For the same study, power analyses for $R^2$ and coefficients usually provide different sample size estimates (generally, the $R^2$ requires the smallest sample size, particularly when there are many predictors). In many cases, researchers are interested in detecting significant effects for coefficients and $R^2$. With those goals in mind, it is important to choose a sample size based on power analyses that reflect all of the effects of interest.

Approaches exist for accurately estimating individual coefficient power for designs with multiple predictors. However, most of these approaches require complex inputs such as partial $R^2$ (G*Power) or variance inflation factors (PiFace). These values require extensive calculations to avoid estimation errors. I expect most researchers would be hard pressed to derive reasonable estimates of these values. Aberson (2010) presents an alternative wherein an approach utilizing only zero-order correlations between variables allows for accurate power analyses for designs with multiple predictors. This approach allows for estimation of multiple forms of power within the same analysis (e.g., power for two coefficients and $R^2$ *model*). Continued development of user-friendly procedures for dealing with complex designs is an important direction for the future.

### A Focus on Designing around Meaningfully Sized Effects

Instead of choosing from small, medium, or large effects or making other arbitrary choices about effect size, I recommend designing power analyses to detect the smallest effect that is practically meaningful. This is often not entirely obvious, particularly in basic research area. For example, when designing an intervention or similar study, the researcher might ask, "how much impact does the intervention need to make to justify the cost? A classic example from Rosenthal (1995) addresses the effectiveness of aspirin therapy on the reduction of heart attacks. In that study, the researchers found a result so compelling that it led them to terminate their project early, as the findings were so clear that people assigned to take aspirin were less

likely to suffer a heart attack than those taking a placebo. The effect size in this study was $r = 0.034$. In the context of this study, what does that effect size mean? This effect is well below the "small" effect criteria for r (0.10). Nonetheless, this size of an effect reflects a 3.4% decrease in heart attacks. This represents a substantial health benefit. Turning to the cost of such an intervention, aspirin is not only widely available but also inexpensive. Turning the example around and thinking about designing for a study of the effectiveness of a similar therapy, the answer to the question regarding how much impact justifies cost would likely be that even a minimal impact would justify the cost of aspirin treatment. This means we would need to budget for a very large sample to detect effects. As a twist on this example, now imagine we were interested in the effectiveness of an expensive medicine to reduce heart attacks (e.g., a cost of $50 per daily dose). Given the cost of the drug, in this case we might require a larger reduction in heart attacks to term the drug "effective." For many research questions, clear benefits of this nature may not be obvious. However, thinking deeply about different potential outcomes in terms of the size of different effects improves decision making in conducting power analyses.

When cost–benefit analyses are not relevant, I suggest extensive investigation of published literature in your specific area or similar areas. This is a better approach than arbitrarily choosing effect size but there are some problems with this strategy. In line with earlier discussions, this approach does not address if effect sizes in previous studies reflect meaningful outcomes but it does give a sense of what researchers doing similar work tend to find. Problematically, published work tends to favor larger effects (i.e., those that were statistically significant). With increasing skepticism over selective exclusion of nonsignificant replications in multistudy papers (e.g., Francis, 2012), it is important to look to effect sizes from published studies with a critical eye. To address these concerns, a good use of this approach would derive estimates from multiple studies conducted by different researchers (e.g., five studies examining similar effects with a range of $d$ from 0.20 to 0.35) and then use the smaller effects for effect size your power analysis.

### Improving Reporting and Interpretation of Power Analysis

Wilkinson and the Task Force for Statistical Inference noted "[b]ecause power computations are most meaningful when done before data are collected and examined, it is important to show how effect size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size (1999, p. 596)." Similarly, the Publication Manual of the American Psychological Association (2010, p. 30)

directs authors to "[s]tate how [the] intended sample size was determined (e.g., analysis of power or precision)." Despite these recommendations, most manuscripts include no information about power analyses. A cursory review of recent literature in psychology finds that only a small proportion of studies addressed power. Researchers are either not conducting power analyses or reviewers and editors are not encouraging them to report on power.

Among studies that do report power analysis for sample size selection, there appear to be substantial problems in terms of the quality of the power analysis. Take this example from a recent article I reviewed. The authors applied a 2 × 2 between-subjects ANOVA design with predictions about significant main effects and an interaction. The method section included the following information about sample size selection: "A sample of 156 participants yields Power = 0.80 for detecting medium sized effects." In my experience, a sentence such as this satisfies most readers. However, there are numerous problems with this declaration. First, why are "medium-"sized effects interesting? Without a clear justification for why a medium-sized effect is meaningful, effect size selection appears arbitrary. Second, what does the "medium effect" refer to? The authors presented predictions for both main effects and an interaction effect (i.e., three unique null hypotheses). Does this power analysis refer to one main effect, both main effects, the interaction, or all of the effects? Because there are multiple effects, greater specificity is necessary. Third, what exactly does a medium effect for an interaction tell us? Most researchers would be unable to interpret such a value in practical terms. Ultimately, what is missing from this analysis is information about why the authors choose a medium effect, the steps they took to making that determination, and an indication of power for the specific effects of interest.

These problems likely result from a handful of factors. Conducting power analyses is very easy given software advances. Simply input a few values and the computer outputs a result. Unfortunately, without a strong focus on choosing values to input, the computed result can be meaningless. From my experiences in editorial roles, when authors present power analyses, there does not appear to be a great deal of critical evaluation of the analyses presented by reviewers.

*Reporting on Power Analysis.* Researchers should detail in their methods section how they determined effect size for power analysis and all values used in estimation. The following example is appropriate for a two-group comparison: To determine sample size requirements, I defined a meaningful test score difference between the two groups as 5 points on a 100-point exam. Five points is meaningful as it corresponds to what is typically a half

of a letter grade. The examination used to test student understanding of materials is taken from previous courses and typically produces a standard deviation of 10 points. These values correspond to $d = 0.50$. The approaches outlined in Aberson (2010) found a sample of 128 participants (64 per group) yields power of 0.80 for detecting differences between the two groups with $\alpha = 0.05$.

*Power Analysis for Detecting Multiple Effects in the Same Study.* Designs with multiple predictors require attention to power for detecting a set of outcomes rather than just power for individual predictors. For example, a researcher conducting a multiple regression analysis with two predictors is often interested in detecting significant regression coefficients for both of the predictors. Common approaches to power analyses for studies with multiple predictors yield an estimate of power for each predictor individually. However, power for an individual predictor is not the same as power for detecting significance on *both* predictors at the same time. Power to detect multiple effects differs considerably from power for individual effects. In most research situations, power to detect multiple effects is considerably lower than the power for individual effects. I term the power to detect all effects in a study as Power (All). As a simple example, imagine flipping two coins. The probability of Coin #1 coming up heads is 0.50. The probability of Coin #2 coming up heads is also 0.50. These values are analogous to the power of each individual predictor. However, what if we were interested in how likely it was to obtain heads on both coin flips? This is analogous to being able to reject both null hypotheses in the same sample. This probability would not be 0.50; it would be lower (0.25 to be precise). This value is analogous to Power(All). Despite the relative simplicity of the concept, the lack of attention to Power(All) may be a primary source of underpowered research in the behavioral sciences (Maxwell, 2004).

The power to detect a set of effects in a study is a product of the power of the individual predictors and the correlation between those predictors. Taking a simple example, if two predictors have Power = 0.80 and are uncorrelated, Power(All) is simply the product of the two power estimates (0.80*0.80 = .64). This means that a study designed to yield 80% power on both predictors has only 64% chance to detect both effects.

Another issue affecting Power(All) is the number of predictors in a study. For example, a study with three predictors that have individual levels of power of 0.80, would (given uncorrelated predictors) produce Power(All) = 0.51 (0.8*0.8*0.8). For four predictors Power(All) would be 0.41.

These calculations become more complex when dealing with correlated predictors. Correlated predictor variables are common in multiple regression

and many multivariate techniques. As a general rule, if the predictors relate to each other in the same manner that they correlate with the DV (e.g., all positively or all negatively correlated) then stronger correlations among predictors reduces Power(All).

At present, detecting Power(All) for designs with correlated predictors is limited as it requires computer-based simulation methods that are often too complex for widespread use. However, on the basis of the patterns discussed, it is reasonable to suggest that when predictors correlate we can get a rough estimate of Power(All). For example, if there are two predictors with power of 0.80 and 0.90, their product (0.80*90) is 0.72. If those predictors correlate in the same manner as with each other as with the criterion, we expect Power(All) to be less than 0.72. Although not ideal, this approach does identify situations where sample size is too small to provide adequate power to detect multiple effects. An important direction for future work is development of user-friendly approaches to determining power for detecting multiple predictors.

### Underappreciated Influences on Power Areas

There are a number of issues that attenuate power through their influence on effect size. By attenuate power I mean that observed (sample) effect sizes will be smaller than population effects (i.e., the effects you obtain are smaller than they should be). As sample effect size goes down, so does power. For continuously scaled variables, imperfect scale reliability is a common cause of attenuated effect sizes (i.e., will make correlations and differences in samples smaller than between the constructs in the population; Hunter & Schmidt, 1994). Reliability refers to the consistency of a measure. The most common estimate of power in the behavioral sciences is internal consistency. Internal consistency addresses how strongly items within a scale relate to each other. The most common estimate of internal consistency is Cronbach's alpha, a measure that ranges from 0–1.0, with 1.0 indicating perfect reliability. As an example of the influence of reliability on power, if two variables correlate in the population at 0.50 but our measures both produce alpha of 0.80, the observed correlation would average 0.40 (a 20% reduction in size). If the variables had lower reliability (alpha = 0.50), the observed correlation would average 0.25. Another issue is artificial dichotomization of continuously scaled variables. Artificial dichotomization involves taking a continuous scaled variable (e.g., an item on a 1–100 scale) and breaking scores into two groups (one above the median, one at or below the median). Dichotomization through this approach produces reduces observed effect size by roughly 20% (Hunter & Schmidt, 1990). Similalry, restriction of range of study variables (i.e., range of values in sample smaller than population)

also reduces observed effects. Finally, violation of test assumption such as homogeneity of variance, homoscedasticity, and sphericity often leads researchers to use tests and adjustments that account for those problems. In general, these approaches focus on reducing the possibility of type I errors by making it more difficult to reject the null hypothesis. These approaches reduce power. If the work in your area tends to suffer from assumption violations, be sure to account for that with increased sample size to offset the loss of power (see Aberson, 2010 for applications). Although there is a good understanding of how these factors impact observed effects, designing power analyses that address these considerations remains an important area for future study.

## REFERENCES

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York, NY: Psychology Press.

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

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*, 585–594. doi:10.1177/1745691612459520

Howell, D. C. (2011). *Fundamental statistics for the behavioral sciences* (7th ed.). Belmont, CA: Duxbury.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*, 334–349. doi:10.1037/0021-9010.75.3.334

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York, NY: Russell Sage.

Kelley, K. (2007). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods*, *39*, 979–984. doi:10.3758/BF03192993

Lenth, R. V. (2000, August). Two sample size practices that I don't recommend. Paper presented at the Joint Statistical Meeting, Indianapolis, IN.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health-psychology related journals. *Health Psychology*, *20*, 76–78. doi:10.1037/0278-6133.20.1.76

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163. doi:10.1037/1082-989X.9.2.147

(2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Rosenthal, R. (1995). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice*, *2*, 133–150. doi:10.1111/j.1468-2850.1995.tb00035.x

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. doi:10.1037/0003-066X.54.8.594

## FURTHER READING

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York, NY: Psychology Press.

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

Kelley, K. (2007). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods*, *39*, 979–984. doi:10.3758/BF03192993

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi:10.1146/annurev.psych.59.103006.093735

## CHRISTOPHER L. ABERSON SHORT BIOGRAPHY

**Christopher L. Aberson** is currently Professor of Psychology at Humboldt State University. He earned his PhD at the Claremont Graduate University in 1999. His topical research interests focus broadly on prejudice and racism and has been published in outlets such as Personality and Social Psychology Review, Group Processes and Intergroup Relations, and European Journal of Social Psychology . He is presently Associate Editor of the Group Processes and Intergroup Relations. His text, *Applied Power Analysis for the Behavioral Sciences* was published in 2010. Chris gives regular workshop presentations focused on statistical methods (primarily power analysis).

Textbook Webpage: http://www.psypress.com/books/details/9781848728356/

Other URLS TO COME – We are currently overhauling our department web pages

## RELATED ESSAYS

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure *(Methods)*, Robert Boruch and Alan Ruby

Hierarchical Models for Causal Effects (*Methods*), Avi Feller and Andrew Gelman

Regression Discontinuity Design (*Methods*), Marc Meredith and Evan Perkoski

Text Analysis *(Methods)*, Carl W. Roberts