Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation

Scott E. Maxwell,[1] Ken Kelley,[2] and Joseph R. Rausch[3]

[1]Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556; email: smaxwell@nd.edu

[2]Inquiry Methodology Program, Indiana University, Bloomington, Indiana 47405; email: kkiii@indiana.edu

[3]Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455; email: rausch@umn.edu

Annu. Rev. Psychol. 2008. 59:537–63

The *Annual Review of Psychology* is online at http://psych.annualreviews.org

This article's doi: 10.1146/annurev.psych.59.103006.093735

0066-4308/08/0203-0537$20.00

Key Words

effect size, confidence intervals, cumulative science

Abstract

This review examines recent advances in sample size planning, not only from the perspective of an individual researcher, but also with regard to the goal of developing cumulative knowledge. Psychologists have traditionally thought of sample size planning in terms of power analysis. Although we review recent advances in power analysis, our main focus is the desirability of achieving accurate parameter estimates, either instead of or in addition to obtaining sufficient power. Accuracy in parameter estimation (AIPE) has taken on increasing importance in light of recent emphasis on effect size estimation and formation of confidence intervals. The review provides an overview of the logic behind sample size planning for AIPE and summarizes recent advances in implementing this approach in designs commonly used in psychological research.

Contents

INTRODUCTION AND OVERVIEW

One of the most frequently asked questions of a statistical consultant is how large a sample is needed for a specific research project. This question is usually couched in terms of designing a study with sufficient statistical power to achieve a statistically significant result. Given recent arguments in favor of reducing the role of null hypothesis significance testing (NHST), such sample size planning might seem less important. In fact, we believe that sample size planning remains a vital aspect of research, regardless of one's position on the NHST controversy. In particular, we argue that sample size planning is important not only for an individual investigator who aspires to publish, but also for a discipline that aspires to create a cumulative science.

From the standpoint of an individual investigator, statistical power is clearly important because most publication outlets in psychology implicitly require statistically significant results as a prerequisite for publication. Thus, investigators who want to publish need to have adequate power. Despite the obvious nature of this statement, literature reviews continue to show that underpowered studies persist, not just in psychology but also in other disciplines (Bezeau & Graves 2001, Cashen & Geiger 2004, Chan & Altman 2005, Maggard et al. 2003). Maxwell (2004) suggests that one reason for their persistence is the simple fact that most studies involve multiple hypothesis tests. Even though the power of any single test may be low by any reasonable standard, the opportunity to conduct multiple tests makes it highly likely that something of interest will emerge as statistically significant. Unfortunately, Maxwell (2004) goes on to show that the consequence for the discipline is an abundance of apparent contradictions in the published literature. Other authors such as Greenwald (1975) and Ioannidis (2005) have similarly shown the importance of power for the development of a cumulative science.

O'Brien & Castelloe (2007) extend this idea through what they define to be "crucial Type I" and "crucial Type II" error rates. The crucial Type I error rate is the probability that the null hypothesis is true when the null hypothesis is rejected. Similarly, the crucial Type II error rate is the probability that the null hypothesis is false when the null hypothesis is not rejected. All too many researchers may be under the false impression that these crucial error rates are simply α and β. In reality, however, as O'Brien & Castelloe (2007) show, these crucial error rates in fact are given by

$$\begin{aligned}\alpha^* &= Prob(H_0\, true | p \le \alpha) \\ &= \frac{\alpha(1-\gamma)}{\alpha(1-\gamma)+(1-\beta)\gamma} \qquad (1) \\ \beta^* &= Prob(H_0\, false | p > \alpha) \\ &= \frac{\beta\gamma}{\beta\gamma+(1-\alpha)(1-\gamma)}, \qquad (2)\end{aligned}$$

where α^* is the crucial Type I error rate, β^* is the crucial Type II error rate, α is the usual Type I error rate, β is the usual Type II error rate, and γ is the prior probability that the null hypothesis is false (or, from a frequentist perspective, the proportion of all relevant studies for which the null hypothesis is false). A key point emphasized by O'Brien & Castelloe (2007) is that all other things being equal, greater power reduces both types of crucial error. As a result, statistical results are more trustworthy when power is high.

Thus, adequate power is an issue not only for an individual investigator who aspires to publish, but also for a discipline that aspires to develop a cumulative literature. The effect on the field may in fact be one reason why old theories in psychology never seem to die, but rather only fade away due to what is claimed to be the slow progress in psychology (Meehl 1978). O'Brien & Castelloe (2007) provide a related perspective by discussing the relation between crucial error rates and the "March of Science."

The concept of power is relevant only in the context of hypothesis testing, because the very definition of power is the probability of rejecting the null hypothesis in favor of an alternative hypothesis when the alternative hypothesis is true. While acknowledging the controversial nature of significance testing (Harlow et al. 1997, Nickerson 2000), we believe that power analysis should play an important role in psychological research. A full treatment of this issue is beyond the scope of this review, so instead we borrow from Jones & Tukey (2000), who among others have pointed out that in many situations a two-tailed hypothesis test provides information about a potentially important question, namely the direction of an effect. In particular, single-degree-of-freedom two-tailed hypothesis tests generally lead to one of three conclusions about a parameter or about a difference between parameters: (*a*) it is negative, (*b*) it is positive, or (*c*) the sign cannot be determined, so it plausibly could be negative, zero, or positive.

How relevant to psychological research is the information provided by hypothesis tests? We submit that sometimes it is of crucial importance, whereas other times it may be a foregone conclusion. For example, consider Festinger's & Carlsmith's (1959) classic study of cognitive dissonance. Would participants rate a boring study more highly if they received a payment of $1 or a payment of $20 (roughly $7 and $140, respectively, in 2006 dollars)? As predicted by cognitive dissonance theory, participants who received $1 rated the study more highly than participants who received $20. How does this relate to sample size planning? We would maintain that the primary goal of this study was to determine the sign of the difference in mean rating between the two participant groups. In particular, which group would produce the higher mean rating could not be predicted with certainty prior to conducting the study. Thus, the hypothesis test allowed the investigators to answer their primary research question. Notice that this question was not literally whether the groups would produce identical ratings, but rather which group would produce the larger rating. This study continues to be a classic at least in part because competing

theories predicted different directions for the difference. Whether the mean difference was small, medium, or large was basically irrelevant. Thus, sample size planning for power should play a critical role here because the goal is to establish the direction of the mean difference.

Now consider a different example. Sternberg & Williams (1997) examined the ability of Graduate Record Examinations (GRE) scores to predict various measures of graduate school success. Here it is difficult to imagine that the direction of the correlation would not be positive. Instead, the question of interest is the magnitude of the correlation. As a result, power takes on reduced importance. However, this hardly makes sample size planning irrelevant, because the size of the sample will directly affect the precision and thus the accuracy with which the population correlation is estimated. For example, a correlation of 0.40 obtained in a sample of 100 yields a 95% confidence interval for the correlation that stretches from 0.22 to 0.55. The fact that the interval excludes zero allows a conclusion that the population correlation is positive, but the magnitude could be anywhere from halfway between small and medium to larger than large according to Cohen's (1988) conventions. If this interval is deemed too wide, the simplest solution (other than decreasing the level of confidence below 95%) is to obtain a larger sample.

Notice the different emphases in the cognitive dissonance and GRE examples. In the first example, sample size should be driven primarily by considerations of power. In the second example, the main goal is to estimate the magnitude of a parameter, which leads to a different approach to sample size planning. In particular, this review describes a variety of procedures for choosing sample size to obtain accurate parameter estimates, in the sense that there is a sufficiently narrow range of plausible values for the parameter of interest, as judged by the width of the corresponding confidence interval. However, we need to be clear that methods of sample size planning for accuracy have only recently begun to be widely developed for many statistical methods. Thus, certain sections of this review focus exclusively on sample size planning for power. We should also add that sample size planning for power is not at all incompatible with sample size planning for accuracy; instead, both perspectives will often be important and need to be considered together because often the goal should be to obtain an accurate estimate of a parameter and also to ascertain whether the parameter is negative, zero, or positive.

Confidence intervals provide a useful organizational framework for simultaneously considering the direction, the magnitude, and the accuracy of an effect. Direction is unambiguous (within the usual limits of probabilistic certainty) when a confidence interval fails to include zero as a plausible value. Thus, from this perspective, power can often be construed in terms of desiring a sufficiently high probability that a confidence interval based on one's observed data will not contain a value of zero. Magnitude requires consideration of precision and accuracy. If estimating the magnitude of a parameter is important, it follows immediately that the width of a confidence interval for this parameter should be considered, along with the center of the interval. A narrow interval results when the standard error of the parameter estimate is small, which is equivalent to saying that the parameter is estimated precisely. Accuracy entails not only precision but also an interval that tends to contain the true population value. In many situations, accuracy and precision go hand in hand, because many estimators are unbiased or at least consistent. Readers interested in additional discussion of the relationship between accuracy and precision can consult Kelley & Maxwell (2003, 2008), Kelley et al. (2003), and Kelley & Rausch (2006).

A CLOSER EXAMINATION OF POWER AND ACCURACY

Consider a researcher who is planning a two-group study where the goal is to compare

mean scores in the treatment and control groups. For simplicity, assume that participants are randomly assigned to groups, with responses independently determined. Further suppose that normality and homogeneity of variance are plausible assumptions, so the researcher plans to analyze these data with an independent groups *t*-test with a two-tailed alpha level of 0.05.

Suppose this researcher desires a power of 0.80. One immediate dilemma is the necessity of specifying an effect size. Suppose the researcher decides to follow Cohen's (1988) guidelines and on this basis specifies a medium effect size (i.e., a population Cohen's *d* of 0.50). The researcher discovers that he or she will need to have 64 participants per group, or a total sample size of 128, assuming no attrition. Now suppose the researcher conducts the study, and it so happens that the standardized sample mean difference between groups turns out to be exactly 0.50, and thus is exactly medium according to Cohen's (1988) conventions. The corresponding *t* value equals 2.83, which is statistically significant at the 0.05 level. This might seem to be a happy ending to the story—the apparent conclusion is that there is a true mean difference between the groups, and the difference corresponds to a medium effect size. However, this effect size value of 0.50 is only an estimate and is itself subject to variability. Recent authoritative sources have recommended that confidence intervals accompany effect size estimates. For example, the *Publication Manual of the American Psychological Association* (Am. Psychol. Assoc. 2001) follows earlier advice offered by Wilkinson et al. (1999) in stating that "The reporting of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results . . . . The use of confidence intervals is therefore strongly recommended" (2001, p. 22). Similarly, the American Educational Research Association reporting standards state that "an indication of the uncertainty" of effect size indices "should be included" (Am. Educ. Res. Assoc. 2006, p. 10).

Heeding the advice of the *Publication Manual of the American Psychological Association* (Am. Psychol. Assoc. 2001) and the *Standards for Reporting on Empirical Social Science Research in AERA Publications* (Am. Educ. Res. Assoc. 2006), our hypothetical researcher proceeds to form a confidence interval. Specifically, a 95% confidence interval for the population value of Cohen's *d* turns out to range from 0.15 to 0.85. Suddenly, it is not at all clear that the true effect here is medium even though the sample value of Cohen's *d* was exactly 0.50. In fact, the confidence interval reveals that the effect could plausibly be smaller than small (i.e., less than 0.20) or larger than large (i.e., greater than 0.80).

Goodman & Berlin (1994) provide a link between power and precision. In particular, they derive the following simple rule-of-thumb approximate relations between confidence intervals and detectable differences:

$$\text{Predicted 95\% CI} = \text{observed difference} \pm 0.7\Delta_{0.80} \quad (3)$$

$$= \text{observed difference} \pm 0.6\Delta_{0.90}, \quad (4)$$

where $\Delta_{0.80}$ = true difference for which there is 80% power and $\Delta_{0.90}$ = true difference for which there is 90% power.

Equation 3 shows why our conscientious hypothetical investigator obtained such a wide confidence interval for Cohen's *d* even while planning a study with adequate power. Recall that the researcher chose a sample size that would provide power of 0.80 for a medium effect size of 0.50. Substituting a value of 0.50 into Equation 3 produces an interval stretching 0.35 below and 0.35 above the observed difference. Because the observed Cohen's *d* was 0.50, the accompanying confidence interval ranges from 0.15 to 0.85. Notice by implication that regardless of the observed effect size, a total sample size of 128 (assuming equal sample sizes of 64 per group) will result in a 95% confidence interval for Cohen's *d* whose total width will be approximately 0.70.

The clear message here is that although a total sample size of 128 may be adequate for power, this sample size does not provide a highly accurate estimate of the population Cohen's *d*. We revisit procedures for designing studies to obtain a sufficiently accurate estimate of Cohen's *d* below.

It is important to emphasize that Equations 3 and 4 provide a useful rule of thumb for sample size planning for any parameter estimate or effect size; the accuracy of the approximation will depend on the extent to which the relevant standard error is independent of the effect size, an issue to which we return below. For example, consider the goal of ascertaining the relation between GRE scores and graduate school success. Suppose the sample size is chosen to be 84 to have power of 0.80 to detect a medium correlation of 0.30 according to Cohen's (1988) conventions. It immediately follows from Equation 3 that the total width of a 95% confidence interval for the population correlation coefficient will be approximately 0.42. For example, if the observed correlation in the sample happens to equal 0.30, the corresponding 95% confidence interval will stretch from 0.09 to 0.48, close to but not literally identical to the width implied by Equation 3 because the standard error of the sample correlation coefficient depends partly on the value of the correlation itself. The confidence interval once again reveals all too clearly, just as it did in the previous example of Cohen's *d*, that considerable uncertainty remains about the true value of the population correlation coefficient.

These examples illustrate that even if sample sizes are sufficiently large to guarantee adequate power, they may not be large enough to guarantee accurate parameter estimates. In reality, these examples probably underestimate the severity of the problem in the current psychological literature because, as mentioned above, literature reviews continue to show that studies tend to be underpowered to detect a medium effect. If studies are adequately powered to detect only a large effect, Equations 3 and 4 show that the ensuing confidence intervals will be wider yet. This underscores Cohen's hypothesis about why confidence intervals tend not to be reported in the literature. In his classic 1994 *American Psychologist* article, he stated, "I suspect that the main reason they are not reported is that they are so embarrassingly large!" (Cohen 1994, p. 1002). However, failing to report confidence intervals simply provides a false sense of certainty, and sets readers up to interpret seemingly discrepant values reported in different studies as being contradictory of one another. Such embarrassment may also reflect an unrealistic expectation about the extent to which a single study can provide a definitive answer, a topic that we discuss below.

The most general point here is that sample size planning should sometimes focus on obtaining a sample large enough to have an adequate probability to reject the null hypothesis, whereas other times the focus should be on an adequate probability of obtaining a sufficiently narrow confidence interval. The sample size necessary to obtain an accurate estimate can be larger than the sample size necessary for adequate power, but the reverse can also be true, depending primarily on the size of effect to be detected. In fact, in many situations it should be important to achieve two goals: (*a*) reject the null hypothesis and establish the direction of an effect, and (*b*) estimate the effect accurately. The first of these implies the need to plan sample size in terms of power, whereas the second implies the need to plan sample size in terms of accuracy. Work that is described below has begun to develop sample size planning methods that accomplish both of these goals simultaneously.

CONCEPTUAL FOUNDATION FOR SAMPLE SIZE PLANNING FOR ACCURACY

Most behavioral researchers realize the importance of sample size planning and power analysis in order to have an appropriate probability of rejecting the null hypothesis when it is false. However, fewer researchers are as

familiar with the role of sample size planning in order to avoid having to present "embarrassingly large" confidence intervals. The current section will provide a general conceptual framework for the specific examples that follow in later sections.

The basic idea of sample size planning for accuracy (i.e., accuracy in parameter estimation, or AIPE) is based on controlling the width of the confidence interval of interest. For example, consider the case of a confidence interval for a difference between two independent means. Assuming normality, homogeneity of variance, and equal group sizes, a 95% confidence interval for a difference between independent means can be written as

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{.975,2n-2}\, s_P \sqrt{2/n}, \qquad (5)$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means for each group, $t_{.975,2n-2}$ is a critical t value corresponding to an alpha level of 0.05 two-tailed with $2n-2$ degrees of freedom, s_P is the pooled sample standard deviation, and n is the sample size per group. Suppose a researcher wants his or her confidence interval to have a "half width" of ω. In other words, the desired ensuing 95% confidence interval will be

$$(\bar{Y}_1 - \bar{Y}_2) \pm \omega. \qquad (6)$$

Notice that the effect of interest here, namely the population mean difference, has a desired precision equal to ω, in the sense that the true population difference should (with probability of 0.95) be within ω units of the sample difference (notice the full confidence interval width is 2ω).

Equations 5 and 6 show that the confidence interval will have the desired width if

$$\omega = t_{.975,2n-2} s_P \sqrt{2/n}. \qquad (7)$$

It might seem that we could easily obtain the necessary sample size simply by solving Equation 7 for n:

$$n = \frac{2t^2_{.975,2n-2} s^2_P}{\omega^2}. \qquad (8)$$

However, three factors prevent Equation 8 from providing the actual desired sample size per group. First, the t value in the numerator depends on n, and thus n is necessarily on both sides of the equation. However, except for very small sample sizes, the z is very close to the t, so introducing a z value of 1.96 for a two-sided 95% confidence interval only very slightly underestimates the actual desired sample size. Second, the variance term in the numerator is a sample statistic. This dilemma can be solved by replacing s^2 with a population variance σ^2. Of course, this leads to other issues, because σ^2 is itself unknown. Nevertheless, the sample size obtained from using σ^2 can be thought of as a conditional sample size, based on a working value of the population variance. Ongoing research addresses the possibility of updating this variance quantity based on early looks at one's data (Coffey & Muller 2003, Proschan 2005). Yet another alternative is to express the desired half-width ω in standard deviation units, in which case σ^2 appears in both the numerator and denominator and thus cancels itself out of the equation.

A third complication is less obvious and pertains specifically to sample size planning for accuracy. Even if a researcher were fortunate enough to use the correct value of σ^2 in the equation for sample size, the actual confidence interval will be based on the sample variance s^2_P, not on σ^2. As a result, even if the correct value of σ^2 is substituted into the expression for sample size, the result will be an interval whose expected width approximately equals the desired width. However, whenever the sample variance happens by chance to be larger than the population variance, Equation 5 shows that the interval obtained from the sample data will be wider than desired. As a result, AIPE requires the specification of "tolerance," which is the probability that the interval will be wider than desired. For example, a researcher might specify that he or she wants to be 80% certain of obtaining an interval no wider than the desired width, in which case tolerance would equal 0.20. Such a goal clearly requires a larger sample size than if the researcher were willing to tolerate only the expected interval width being sufficiently

narrow. Sections of the review below provide references for incorporating this tolerance value into sample size planning for various effects.

SPECIFICATION OF EFFECT SIZE

One of the most troublesome aspects of sample size planning is the necessity to specify an effect size. In fact, as Lipsey (1990, p. 47) puts it in his chapter entitled "Effect Size: The Problematic Parameter," "The problem that is perhaps most responsible for inhibiting statistical power analysis in the design of treatment effectiveness research, however, is the fact that the effect size is generally both unknown and difficult to guess." Senn (2002, p. 1304) addresses this criticism of power analysis by pointing out, "The difference you are seeking is not the same as the difference you expect to find, and again you do not have to know what the treatment will do to find a figure. This is common to all science. An astronomer does not know the magnitude of new stars until he has found them, but the magnitude of star he is looking for determines how much he has to spend on a telescope." Also important is the point that power is not literally a single number but instead is a function defined over parameter values consistent with the alternative hypothesis. As such, power curves and response surfaces show how power changes as a function of such factors as effect size and sample size and thereby provide much more information than a single number.

Adding to the confusion is considerable disagreement about what magnitude of effect is truly important. McCartney & Rosenthal (2000) and Prentice & Miller (1992), among others, have argued that psychologists tend not to realize that effects conventionally thought of as small or even less than small may in fact be very important, either scientifically or practically. Unfortunately, in practice, sample size planning often is based on exactly the opposite perspective, whereby power becomes adequate to detect only large effects. As Goodman & Berlin (1994, p. 203) state, "A typical sample size consultation often resembles a ritualistic dance. The investigator usually knows how many participants can be recruited and wants the statistician to justify this sample size by calculating the difference that is 'detectable' for a given number of participants rather than the reverse . . . . The 'detectable difference' that is calculated is typically larger than most investigators would consider important or even likely." This description makes it abundantly clear why some researchers may view the sample size planning process as anything but scientific.

In principle, it would seem that researchers who design studies with sufficient power to detect only large effects would end up only hurting themselves, because unless they obtain statistically significant results, they may be unlikely to publish their results. However, any such self-correcting mechanism is likely to operate very gently if at all because almost all studies involve multiple hypothesis tests. As Kelley et al. (2003) and Maxwell (2004) point out, even if the power of any single test is low, the power to detect some effect among multiple tests can easily be quite high. In this sense, the system provides little direct incentive for researchers to adopt a procedure whereby they choose sample size based on a serious consideration of an effect size.

The discipline pays a price for underpowered studies even if individual researchers may not. First, as we have already mentioned, underpowered studies tend to produce a literature with apparent contradictions. Second, as Goodman & Berlin (1994), Hunter & Schmidt (2004), and Maxwell (2004) have pointed out, such apparent contradictions may in fact reflect nothing more than inherent sampling variability. Third, reporting results only as either significant or nonsignificant exacerbates the problem. Much better would be to report results in terms of confidence intervals because they display the uncertainty in effects, thus preventing readers

from overinterpreting the presence of multiple asterisks next to small *p*-values.

A major advantage of sample size planning for accuracy is that sample size formulas for narrow confidence intervals can be much less dependent on the actual value of the population parameter in question than are sample size formulas for power. For example, the mean difference between groups does not appear in Equation 8, so sample size planning here is independent of effect size. In some situations (described below), the desired sample size is not independent of the underlying effect size. However, in such situations sample size for accuracy is often less dependent on the parameter value than is sample size for power. As a result, sample size planning from the AIPE perspective frequently overcomes Lipsey's (1990) major stumbling block for sample size planning because the unknown effect size may be relatively unimportant.

SAMPLE SIZE PLANNING FOR SPECIFIC DESIGNS AND ANALYSES

Comparing Two Independent Groups

The comparison of means via the two-group *t*-test is common in psychological research. As with most sample size planning procedures, obtaining sufficient power has dominated sample size planning for comparing two independent groups. Since Cohen's (1962) cataloging of typical standardized effect sizes in abnormal-social psychological research, researchers have often used Cohen's rules of thumb regarding small, medium, and large effects sizes for the standardized mean difference when planning sample size and interpreting study results.

Of course, rejecting a null hypothesis concerning mean differences may not be as informative as forming a confidence interval for the mean difference or Cohen's *d*. Thus, the AIPE approach to sample size planning is more appropriate than the power analytic approach in some situations. Due to the distributional differences between the unstandardized and standardized mean difference, AIPE sample size planning is not equivalent for these two effect sizes. Kelley et al. (2003) discuss sample size planning from an AIPE perspective in the context of two groups for the unstandardized mean difference (see also Beal 1989). Kelley & Rausch (2006) develop sample size planning procedures from the AIPE perspective for the population standardized mean difference. Both Kelley et al. (2003) and Kelley & Rausch (2006) also compare the power analytic and AIPE approaches to one another. In the context of AIPE for the unstandardized mean difference, the width of the confidence interval is independent of the size of the mean difference (recall Equation 5), which implies that the only parameter specification required is the common variance (Kelley et al. 2003). Necessary sample size for the standardized mean difference is not independent of the population standardized mean difference, but in practice, it depends relatively little on the size of the effect (Kelley & Rausch 2006).

Figure 1 shows necessary sample size per group as a function of the size of the population standardized mean difference (between 0.10 and 0.50) for power of 0.50, 0.80, and 0.95 (where $\alpha = 0.05$, two-tailed) and for desired 95% confidence interval widths of 0.35, 0.25, and 0.15. The smaller the value of the population standardized mean difference, the larger the sample size for a specified level of power. However, the larger the standardized mean difference, the larger the sample size for a specified confidence interval width, albeit the increase in sample size for larger standardized mean differences is minimal. Thus, necessary AIPE sample size for Cohen's *d* depends almost entirely on the desired confidence interval width. Such a realization should help to ease qualms about Lipsey's (1990) "problematic" unknowable effect size parameter in this situation.

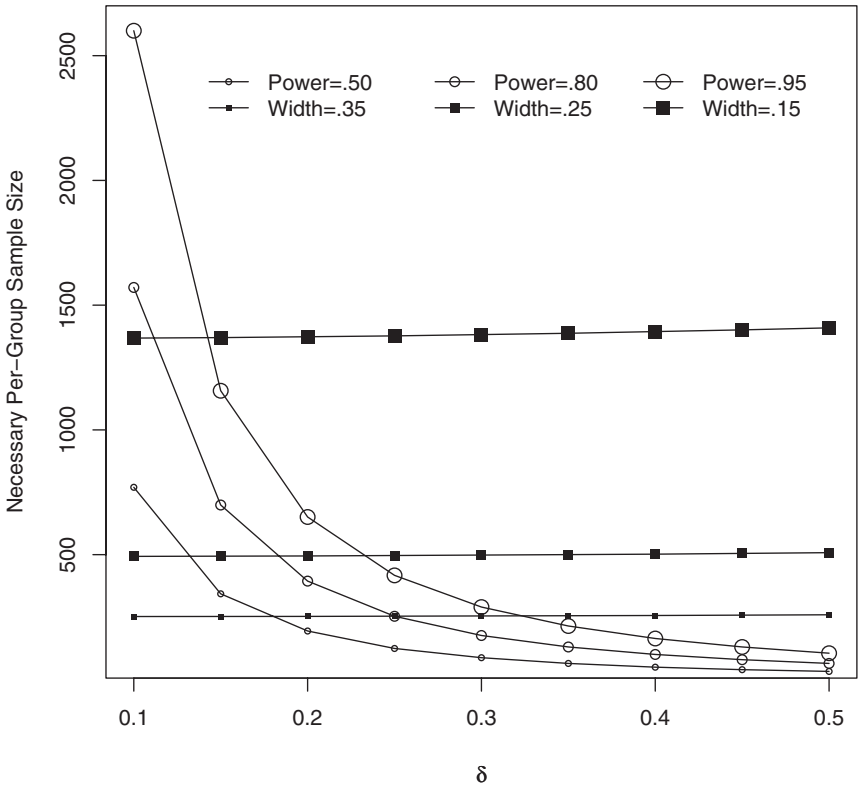

Figure 1

Necessary per-group sample size as a function of effect size for desired power and desired confidence interval width. Adapted from Kelley & Rausch (2006), with permission.

Adjustment for Multiple Comparison Procedures

In the analysis of variance (ANOVA) framework, multiple comparisons are commonly performed to address targeted questions about mean differences. Such contrasts are generally evaluated with a modified critical value due to the effect of multiplicity on the Type I error rate. Miller (1981) and Hsu (1996) provide comprehensive reviews on issues surrounding multiple comparisons. When a multiple comparison procedure will be used in data analysis, sample size planning should take this into account. Without such consideration, sample size will likely be too small.

Pan & Kupper (1999) develop methods for planning sample size from the AIPE perspective (both for the expected width and desired tolerance) for multiple comparisons, and Hsu (1989, 1996) provides an analogous discussion from the power analytic perspective. Both Pan & Kupper (1999) and Hsu (1989, 1996) develop methods for the most commonly used multiple comparison procedures (e.g., Bonferroni, Tukey, Scheffé, and Dunnett). Williams et al. (1999) discuss alternative multiple comparison procedures, including a sequential procedure for controlling the false discovery rate (FDR; Benjamini & Hochberg 1995). Although controlling the FDR tends to yield more power than controlling the familywise error rate, as of yet no formal sample size planning procedures currently exist for the FDR.

Multiple Regression

Kelley & Maxwell (2008) discuss sample size planning for multiple regression in a two-by-two framework, where one dimension represents the goal of the study, power or accuracy, and the other represents the effect size of interest, omnibus or targeted. Operationally, in

multiple regression the omnibus effect is the squared multiple correlation coefficient, and the targeted effect is a specific regression coefficient. Thus, the way in which sample size is planned, and indeed the sample size itself, should depend on the question of interest.

Cohen (1988) details sample size planning for desired power for the omnibus effect (i.e., the squared multiple correlation coefficient) and a targeted effect (i.e., a particular regression coefficient). Commonly cited rules of thumb that pervade the literature on sample size for multiple regression are rarely appropriate (Green 1991, Maxwell 2000). Maxwell (2000) develops a set of procedures to plan sample size for targeted effects (see also Cohen 1988). Something not obvious is the fact that it is entirely possible for each regression coefficient to require a different sample size in order for there to be the same degree of power. More importantly, and not previously well documented in the literature, is that an appropriate sample size for the test of the squared multiple correlation coefficient in no way implies an appropriate sample size for the test of a particular regression coefficient.

Kelley & Maxwell (2003), later updated and generalized in Kelley & Maxwell (2008), develop methods for planning sample size for a targeted regression coefficient from the AIPE perspective. Much like the discussion regarding unstandardized and standardized mean differences, the value of an unstandardized regression coefficient is independent of confidence interval width, yet for standardized regression coefficients, a relatively small relation exists between the size of the standardized regression coefficient and the width of the confidence interval (Kelley & Maxwell 2008). This demonstrates that the AIPE approach to sample size planning for a targeted regression coefficient is easier to implement than the corresponding power analysis in the sense that the appropriate sample size is much less sensitive to the unknown effect size.

Kelley & Maxwell (2008) and Kelley (2007) develop AIPE sample size planning methods for the squared multiple correlation coefficient for fixed and random regressors, respectively. The necessary AIPE sample size for the squared multiple correlation coefficient, contrary to a targeted regression coefficient, depends heavily on the value of the effect size. Algina & Olejnik (2000) develop a related method of planning sample size so that the observed squared multiple correlation coefficient is within a specified distance from the population value with some desired probability. Algina and colleagues have also developed sample size procedures for other important effects not often considered when planning sample size for multiple regression. In a cross-validation context, Algina & Keselman (2000) present methods to ensure that the squared cross-validity coefficient is sufficiently close to its upper limit, the squared population multiple correlation coefficient. Along the same lines as Algina & Olejnik (2000), Algina & Olejnik (2003) develop sample size planning procedures for the zero-order, the zero-order squared, and the partial correlation coefficient. Algina et al. (2002) also apply the sample size approach of Algina & Olejnik (2000) to the difference between squared multiple correlation coefficients (i.e., the squared semipartial correlation coefficient) in nested models.

The General Linear Multivariate Model

Reviews of sample size planning for the multivariate general linear model have been provided by Muller et al. (1992) and O'Brien & Muller (1993). A notable development in sample size planning for the general linear multivariate model is the work of Jiroutek et al. (2003). Their work combines the power analytic approach with an approach similar to AIPE, where the goal is to obtain a narrow confidence interval, conditional on the population value being contained within the observed interval. Thus, the approach of Jiroutek et al. (2003) accomplishes three things simultaneously with a single method: (*a*) an estimate that leads to a rejection of the null hypothesis, (*b*) a corresponding

confidence interval that is sufficiently narrow, and (*c*) a confidence interval that correctly brackets the population parameter, all with some specified probability. The fundamental idea of this approach is to formalize sample size planning to ensure a specified probability for obtaining an estimate that is simultaneously accurate and statistically significant. Thus, the work of Jiroutek et al. (2003) is especially valuable for researchers whose goals include establishing both the direction and magnitude of an effect.

Most methods of sample size planning for the general linear model assume fixed predictors. Many psychological predictor variables are continuous, and most continuous variables in psychology are random instead of fixed. Glueck & Muller (2003) review the limited availability of sample size planning methods with random predictors. They also discuss the ramifications of incorporating a random baseline covariate for the calculation of sample size and power. Their methods extend directly to the context of generalized estimation equations (e.g., Liang & Zeger 1986), where not only are discrete and continuous outcomes possible, so too is a flexible correlational structure.

Exploratory Factor Analysis

Various suggestions and rules of thumb for sample size planning permeate the literature on exploratory factor analysis. Many rules of thumb stipulate a desired ratio of sample size to the number of factors, variables, or free parameters. MacCallum et al. (1999, 2001), Hogarty et al. (2005), Nasser & Wisenbaker (2001), and Velicer & Fava (1998) each review the existing literature and show that, in general, such rules of thumb regarding necessary sample size are oversimplified and should not be trusted.

Monte Carlo simulations have clearly shown that necessary sample size depends to a large extent on the goals of the researcher, and planning sample size cannot generally be reduced to rules of thumb. Basing sample size on a ratio relative to the number of variables or an absolute sample size ignores communalities, which greatly affect necessary sample size. MacCallum et al. (1999) develop an approach that explicitly considers communalities as a necessary part of the procedure, with the goal of obtaining more valid factor analysis solutions. MacCallum et al. (2001) extend this work by allowing the factor model to be misspecified.

Confirmatory Factor Analysis and Structural Equation Modeling

Confirmatory factor analysis (CFA) and structural equation modeling (SEM) have become indispensable in much psychological research. SEM and CFA generally evaluate the overall model with chi-square likelihood ratio tests and/or with fit indices. The chi-square likelihood ratio test evaluates exact fit, whereas fit indices quantify how well a model fits the data. One can also consider sample size for specific path coefficients (i.e., targeted effects). Also, instead of considering only power, the AIPE approach could also be used for model fit and targeted effects. Currently, however, AIPE has not been developed in this context. Such a limitation is one that can certainly benefit from future research. As in multiple regression, the way in which sample size is planned should depend on the particular research goals (power and/or AIPE for omnibus and/or targeted effects).

Satorra & Saris (1985) provide an early approach to sample size planning based on the chi-square likelihood ratio test, where a specific but incorrect null model is hypothesized and the noncentrality parameter is determined based on the correct alternative model in order to calculate the probability of rejecting the specified null model.

Muthén & Muthén (2002), Mooijaart (2003), and Yuan & Hayashi (2003) all extend Satorra & Saris (1985) by developing computational approaches to sample size planning, such that data are generated given a specific set of parameters from a specified model in

order to determine power empirically. As Kim (2005) points out, such approaches are based on an alternative model being correctly specified, where all of the model parameters are explicitly stated. Due to the sheer number of parameters of many models, specification of the set of parameters generally proves to be quite difficult.

Rather than approaching sample size from an exact fit perspective, MacCallum et al. (1996) develop sample size planning methods by defining the null hypothesis to be a particular value (generally not zero and thus not a perfect fit) of the root mean square error of approximation (RMSEA; Browne & Cudeck 1993, Steiger 1990, Steiger & Lind 1980). The idea is not necessarily to test an exact model, but rather to determine sample size so that not-good-fitting models could be rejected. This approach is implemented by determining sample size so that the upper limit of a confidence interval for the population RMSEA, given the hypothesized RMSEA value, is less than what is operationally defined as a not-good-fitting model. Such an approach overcomes the problem with the likelihood ratio test that very large samples will essentially always reject the null hypothesis, even for models that are useful (Browne & Cudeck 1993). In this framework, unlike the approach of Satorra & Saris (1985) where a specific null model is specified, the relationship between fit indices and the noncentrality parameter from a noncentral chi-square distribution is exploited so that the fit index itself is specified instead of a large number of individual parameters. For example, given the model-specified degrees of freedom and a hypothesized value of the population RMSEA equal to 0.05, sample size can be planned so that 0.08 is excluded from the confidence interval for the population RMSEA with some specified probability.

MacCallum & Hong (1997) and Kim (2005) extend the methods of MacCallum et al. (1996) to commonly used fit indices other than the RMSEA. MacCallum et al. (2006) further extend the methods of MacCallum et al. (1996) so that differences between the fit of competing models can be tested. Because different fit indices can be used, necessary sample size depends in part on the particular fit index chosen. Hancock & Freeman (2001) provide a tutorial for applied researchers on using the MacCallum et al. (1996) approach, and Hancock (2006) provides a tutorial chapter on general sample size issues in SEM with CFA as a special case. The effect of missing data (e.g., Dolan et al. 2005, Muthén & Muthén 2002) and type of manifest variable (continuous versus discrete; Lei & Dunbar 2004) on power has also been considered.

Longitudinal Data Analysis

Latent growth curve (LGC) models (Bollen & Curran 2006; McArdle 1988; McArdle & Epstein 1987; Meredith & Tisak 1984, 1990) and multilevel models for longitudinal data (Raudenbush & Bryk 2002, Singer & Willett 2003) have become increasingly popular methods for analyzing change. A number of recent approaches have been developed to calculate power and sample size for these models, with continuous or discrete outcomes.

For continuous outcomes, Muthén & Curran (1997) provide an extensive treatment of using LGC models for the analysis of randomized trials and illustrate an approach to power analysis in this context. Hedeker et al. (1999) provide a general framework for sample size planning from a power perspective when designing longitudinal studies to detect group differences, focusing on a mixed/multilevel model approach. Hedeker et al. (1999) also allow for differing degrees and patterns of attrition to be specified for the purpose of examining the effect of missing data on power in longitudinal studies. Similar to the work of Hedeker et al. (1999), Jung & Ahn (2003) provide sample size expressions for sufficient power for group comparisons in longitudinal data analysis that also allow for varying degrees and patterns of attrition. An important difference in these two approaches,

however, lies in Jung & Ahn's (2003) choice of generalized estimating equations (GEE) for the derivation of their results. Winkens et al. (2006) provide expressions and an illustration of the effects of increasing the number of participants per group and the number of measurement occasions on necessary sample size and power for group comparisons in longitudinal data analysis. Winkens et al. (2006) differ from other researchers in this area, however, in that these authors also explicitly incorporate a cost function into their expressions to directly weigh the cost of adding participants versus time points.

Raudenbush & Xiao-Feng (2001) provide expressions for calculating power on group differences in orthogonal polynomial growth model parameters as a function of group sample size, study duration, and the number of measurement occasions. Also in the context of polynomial growth models, Biesanz et al. (2004) note that recoding time generally leads to "a change in the question asked" (p. 43) with respect to the lower-order polynomial growth parameters. Thus, changes in the corresponding power functions due to recoding time are generally due to changes in the meanings of these lower-order growth model parameters. Yan & Su (2006) provide methods for sample size calculation for sufficient power for group differences in longitudinal studies, but also allow nonlinear models (e.g., Bates & Watts 1988) for the growth functions of the two groups.

Other studies have contributed to sample size and power analysis for longitudinal studies of discrete outcome variables. For example, Rochon (1998) proposes a general approach to calculating minimum sample size for power analysis in repeated measures designs based on GEE, where the outcome variable can be discrete or continuous. Rochon (1998) also provides illustrative examples based on binary and Poisson outcome variables. Leon (2004) uses previous work of Diggle et al. (2002) to provide sample size tables for power to detect a treatment effect between two groups as a function of Type I error, the number of repeated binary observations, the group response rates, and the intraclass correlation. Jung & Ahn (2005) derive sample size expressions for sufficient power when comparing group differences in rates of change on a binary variable. Their approach is based on GEE, and it allows researchers to specify the degree and certain patterns of attrition.

Generalized Linear Models

The generalized linear model represents a well-known class of statistical methods that are useful for modeling categorical variables and contingency tables, and more generally, variables that are not normally distributed. A number of sample size planning methods have recently been proposed for this modeling approach, all from a power analysis perspective. Shieh (2005) presents an approach to power and sample size calculation for generalized linear models, which allows researchers to test multiple parameters simultaneously using a Wald test, among other extensions. Lyles et al. (2007) and Newson (2004) provide general, practical approaches for power calculations in the context of the generalized linear model.

The logistic regression model, which can be conceptualized as a special case of the generalized linear model, is used to model categorical or ordinal outcome variables. Recent articles on sample size and power for logistic regression have provided a variety of perspectives on power issues. Tsonaka et al. (2006) describe sample size and power calculations for discrete bounded outcome variables in a randomized trial. Taylor et al. (2006) illustrate the loss of power and necessary increase in sample size to achieve the same level of power when a continuous variable is categorized, either for the purpose of simplifying the analysis or via the process of measuring a continuous variable with ordinal categories. These authors also illustrate the potential utility in utilizing logistic or probit ordinal regression models to minimize the loss of efficiency in such situations.

Vergouwe et al. (2005) illustrate sample size planning for externally validating various logistic regression models using data from patients with metastatic testicular cancer. They argue that this approach is especially useful for model/variable selection. Furthermore, Vaeth & Skovlund (2004) and Hsieh et al. (1998) provide simple approaches to sample size calculations for power in logistic regression. In particular, Hsieh et al.'s (1998) simplified calculations are based on comparisons of means or proportions with a modification based on a variance inflation factor. Also, Strickland & Lu (2003) provide sample size calculations for comparing two groups in pre-post designs when the outcome variable is categorical or ordinal.

Whereas some authors have focused directly on the role of sample size in increasing power in the context of logistic regression, others have focused upon the utility of collecting pretreatment covariates in randomized studies for attaining increases in power. For example, Hernandez et al. (2004) illustrate the potential gain in efficiency either through smaller required sample sizes or through increased power in randomized studies of dichotomous outcomes by incorporating a baseline covariate into the logistic regression model. In fact, Hernandez et al. (2004) report up to a 46% reduction in required sample size for a prespecified level of power for the treatment effect through incorporating a baseline covariate into the analysis.

Cluster Randomized Trials

Cluster randomized trials are often used when it is more practical or feasible to randomly assign groups (i.e., clusters of individuals), as opposed to individual participants, to various treatment conditions (e.g., randomly assigning classrooms instead of students). Sample size planning may be especially important in such situations because even if there are a large number of participants per cluster, power and accuracy will suffer if the number of clusters is small. Thus, researchers should consider a number of alternatives in designing cluster randomized trials. Raudenbush (1997) uses a multilevel modeling framework to evaluate a number of key variables, including the study cost, the number of participants within cluster, the number of clusters, and the increase in statistical efficiency that can be attained by incorporating a pretreatment covariate into the statistical analysis.

Campbell et al. (2004) present a sample size–calculating tool that can be used to determine the necessary number of clusters and participants within cluster to detect a minimally meaningful treatment effect. Moerbeek (2006) evaluates the cost of two approaches for increasing power in a cluster randomized trial: increasing the number of clusters and incorporating pretreatment covariates into the statistical model. In particular, Moerbeek derives expressions that researchers can use to weigh these two alternatives against one another in terms of their relative costs. While also focusing one aspect of their study on the utility of a pretreatment covariate in cluster randomized trials, Murray et al. (2006) compare mixed ANOVA/ANCOVA models to multilevel models with respect to the power to detect an effect and conclude that the mixed model ANCOVA with a pretreatment covariate is preferable.

Federov & Jones (2005) provide a general exposition of a variety of issues in the analysis of cluster randomized trials. In particular, they express a preference for a random-effects model for the analysis of cluster randomized trials and also emphasize the importance of accounting for a number of important variables when designing a cluster randomized trial. Finally, Kraemer & Robinson (2005) clearly describe a number of important methodological issues that need to be carefully considered when designing, conducting, and analyzing cluster randomized trials.

Survival Analysis

Survival analysis is often the method of choice when the outcome variable of interest is the

duration of time until a particular event occurs (e.g., Singer & Willett 2003). Recent work on study design for sufficient power in survival analysis includes Maki (2006), who presents expressions for power and sample size when using a form of the Weibull model to represent hazard functions. In contrast, Vaeth & Skovlund (2004) provide an approach to sample size and power calculations based on the Cox regression model. Furthermore, Schulgen et al. (2005) and Bernardo & Harrington (2001) introduce formulations necessary for sufficient power when comparing two groups in survival analysis, whereas Barthel et al. (2006) present a general approach to power in survival analysis.

Mixture Modeling

The general goal of mixture modeling is to decompose an observed distribution into multiple unobserved distributions. The observed distribution is often termed a composite, as the model implies it is the sum of component distributions. One goal is simply to model an observed, generally nonnormal, distribution. Another goal is to decompose a nonnormal distribution into multiple component distributions, where it is believed the components represent unobserved (i.e., latent) classes/groups. In so doing, the grouping variable can be regarded as missing with the goal of the mixture model then being the recovery of the parameters from each class and/or classification of individuals into the class to which they belong. McLachlan & Peel (2000) provide a survey of mixture models.

Mixture models have been extended to many models, such as regression, confirmatory factor analysis, structural equation modeling, and longitudinal models, among others. Sample size planning for mixture models is thus specific to the particular type of mixture analysis of interest. Furthermore, sample size planning can be based on several different goals: distinguishing between different competing models, assigning each individual to the appropriate class, AIPE for parameter estimates, etc. Thus, crossing the different types of mixture models with the different goals leads to a large set of possible ways to plan an appropriate sample size. Due to the rich questions that mixture models can address, sample size planning in mixture modeling certainly deserves increased attention.

Muñoz & Acuña (1999) and Zheng & Frey (2004) evaluate different combinations of parameters and sample size on the effectiveness of parameter recovery from mixed distributions. Not surprisingly, the consensus is that results are better when sample size is larger. Lubke & Neale (2006) evaluate the role sample size plays in factor mixture models when comparing two-class single-factor mixture models with single-class two-factor models. Again, the general conclusion is that the correct model tends to be selected with higher probability for larger sample sizes and larger class separations. Not obvious, however, is the fact that larger sample sizes are also associated with overestimation of the number of classes. Such a finding supports a common recommendation in the mixture modeling literature that theory should play a role in determining the number of latent classes.

EQUIVALENCE, NONINFERIORITY, AND THE GOOD ENOUGH PRINCIPLE

Equivalence is commonly studied in the medical sciences where two treatments, often drugs, are evaluated to investigate if there is no meaningful clinical difference between them. Evaluation of noninferiority is related to equivalence, except instead of implying there is no meaningful difference, noninferiority implies that one treatment is no worse than the other(s). Tryon (2001) provides a review of equivalence with connection to the psychological literature and the controversy that sometimes surrounds null hypothesis testing. Basic sample size issues for equivalence and noninferiority are discussed in Julious (2004). Liu et al. (2002),

Tang et al. (2002), and Chan (2002) discuss methods for equivalence/noninferiority for binary data, the ratio of proportions in matched-pair designs, and the difference between two proportions, respectively.

Such issues are related to the "good enough" principle and the "good enough belt," where the limits of the belt define what is considered a nontrivial effect (Serlin & Lapsley 1985, 1993). The AIPE approach to sample size planning can be helpful in this context, because if the upper and lower confidence limits are contained within the good enough belt, then evidence exists that meaningful differences are implausible at the specified confidence level. Because AIPE has not yet been developed for all important statistical methods, there is a corresponding deficiency in the sample size planning literature with regard to the good enough principle.

SIMULATION-BASED APPROACHES TO SAMPLE SIZE PLANNING

As has been detailed, sample size planning procedures have been developed for a wide variety of statistical tests. However, these procedures are typically based on standard techniques when all assumptions have been met. Sample size planning procedures for nonstandard analyses (e.g., classification and regression trees) and/or computationally based techniques (e.g., the bootstrap approach to statistical inference) have not generally been developed. Even when sample size planning for power has been developed, at this time there are often no corresponding methods for AIPE. However, a general principle of sample size planning appears to hold: Sample size can be planned for any research goal, on any statistical technique, in any situation with an a priori Monte Carlo simulation study.

An a priori Monte Carlo simulation study for planning an appropriate sample size involves generating random data from the population of interest (e.g., the appropriate parameters, distributional form), implementing the particular statistical technique, and repeating a large number of times (e.g., 10,000) with different sample sizes until the minimum sample size is found where the particular goal is accomplished (e.g., 90% power, expected confidence interval width of 0.15, 85% power and a 1% percent tolerance that the confidence interval is sufficiently narrow). Conducting such an a priori Monte Carlo simulation to plan sample size requires knowledge of the distributional form and population parameters, but this is also true with traditional analytic methods of sample size planning (where normality of the errors is almost always assumed). Muthén and Muthén (2002) discuss sample size planning for CFA and SEM via an a priori Monte Carlo simulation study. In the context of Bayesian inference, M'Lan et al. (2006) and Wang & Gelfand (2002) discuss similar a priori Monte Carlo simulation approaches for AIPE and power.

PLANNED AND POST HOC POWER ANALYSES

Although the fifth edition of the *Publication Manual of the American Psychological Association* (Am. Psychol. Assoc. 2001) encourages researchers to take power considerations seriously, it does not distinguish between planned and post hoc power analysis (also called "observed power" or "retrospective power" in some sources). Hoenig & Heisey (2001) cite 19 journals across a variety of disciplines advocating post hoc power analysis to interpret the results of a nonsignificant hypothesis test. It is important to clarify that post hoc power relies on using the sample effect size observed in the study to calculate power, instead of using an a priori effect size.

Hoenig & Heisey (2001) argue convincingly that many researchers have misunderstood post hoc power. Specifically, a low value of post hoc power does not necessarily imply that the study was underpowered, because it may simply reflect a small observed sample effect size. Yet another limitation of post

hoc power is that Yuan & Maxwell (2005) have shown that post hoc power does not necessarily provide a good estimate of the actual population power even in large samples. Taken together, these perspectives show that confidence intervals and equivalence tests are superior to post hoc power as methods for interpreting the magnitude of statistically nonsignificant effect sizes, whether the goal is to assess support for a trivially small effect size, such as for assessing equivalence or noninferiority, or the goal is to argue that a study was underpowered.

METHODS TO INCREASE POWER AND ACCURACY

The emphasis placed on sample size in most discussions of power and accuracy may lead researchers to conclude that the only factor under their control that can influence power and accuracy is in fact sample size. In reality, although sample size clearly plays a vital role, there are often many other factors under an investigator's control that can increase power and accuracy. In the specific case of a linear model with m predictors, McClelland (2000) notes that the confidence interval for a regression coefficient can be expressed as

$$b \pm t_{N-m-1;\alpha}\sqrt{\frac{MSE}{NV_X(1-R_X^2)}}, \qquad (9)$$

where b is the estimated coefficient, $t_{N-m-1;\alpha}$ is a critical value, N is sample size, MSE is the model mean square error, V_X is the variance of the predictor variable, and R_X^2 is the proportion of variance in the predictor shared with other predictor variables in the model. McClelland describes a variety of possible methods for decreasing the width of the confidence interval, and thereby increasing power, in addition to simply increasing sample size. More generally, Shadish et al. (2002) and West et al. (2000) provide explanations of a number of factors that researchers should consider in their efforts to increase power and accuracy.

META-ANALYSIS AND STUDY REGISTRIES

Researchers who engage in appropriate methods of sample size planning may quickly discover that the sample size needed to obtain adequate power or accuracy exceeds their resources. Investigators who need only determine that effect sizes expected to be large are in fact nonzero in the expected direction may be perfectly able to continue designing studies with relatively modest sample sizes. However, investigators who need to detect small effects or who need to obtain accurate parameter estimates will typically need quite large samples. If psychology is to take seriously the mission to estimate magnitude of effects accurately, researchers may be shocked at how large their samples will need to be.

For example, the standard error of the sample correlation depends on the value of the population correlation coefficient, but for small values of the population correlation, the standard error is approximately $\sqrt{1/n}$. Suppose a researcher wants to pinpoint the true population value of the correlation coefficient to within ± 0.05. A 95% confidence interval for the correlation needs to be based on roughly 1500 cases in order for the interval to have an expected half-width of 0.05 unless the correlation itself is sizable [a large correlation of 0.50 according to Cohen's (1988) conventions would still require more than 850 cases]. Experimentalists do not get the last laugh, because the sample size necessary to obtain a 95% confidence interval with a half-width of 0.05 for a standardized mean difference between two independent means is more than 3000 per group. Sample sizes of such magnitudes presumably explain Hunter & Schmidt's (2004, p. 14) statements that "for correlational studies, 'small sample size' includes all studies with less than a thousand persons and often extends above that" and "for experimental studies, 'small sample size' begins with 3000 and often extends well beyond that."

One way out of this conundrum is to decide that intervals do not need to have

half-widths as narrow as 0.05 to be regarded as sufficiently precise. Other considerations include the same factors besides sample size that can increase precision, such as using a within-subjects design or incorporating covariates in the analysis. Even so, the fact remains that for many types of research programs, very large samples will be required to estimate effects with any reasonable degree of precision, and it will thus generally be difficult to obtain sufficient resources to obtain accurate parameter estimates in a single study.

Meta-analysis provides one potential solution to the lack of precision often observed in individual studies. Cohn & Becker (2003) point out that meta-analysis typically reduces the standard error of the estimated effect size, and thus leads to narrower confidence intervals and therefore more precision. In addition, power is often increased.

Hedges & Pigott (2001, 2004) argue for the importance of conducting power analyses before investing resources in a meta-analysis. They show that standard tests performed as part of meta-analyses do not necessarily have high statistical power, especially tests of heterogeneity of effect sizes, reinforcing the need to conduct a power analysis prior to undertaking a meta-analysis. These two articles together demonstrate how to conduct a power analysis for a variety of tests that might be of interest in a meta-analysis.

Although a major goal of meta-analysis is often to increase power and accuracy, resultant power and accuracy in meta-analysis "can be highly dependent on the statistical model used to meta-analyze the data" (Sutton et al. 2007). In fact, Hedges & Pigott (2001, p. 216) state that "The inclusion in meta-analysis of studies with very small sample sizes may have a paradoxical effect of decreasing the power of random-effects tests of the mean effect size." Along related lines, Lau et al. (1992) suggest that meta-analysis can be used to summarize the state of knowledge at each stage of research. Sutton et al. (2007) implicitly adopt this perspective and thereby argue that sample size planning should often be done not from a perspective of designing a single study with sufficient power, but instead should be done in the context of designing a new study to contribute to a larger body of literature in such a way that an ensuing meta-analysis adding the new study to the extant literature will have sufficient power. They then proceed to present a simulation approach to sample size planning based on this idea. One important result they demonstrate is that in a random effects meta-analytic model, multiple smaller studies can sometimes provide much more power than a single larger study with the same total sample size. This result converges with cautions offered by Schmidt (1996) and Wilson & Lipsey (2001) regarding the hazards of overinterpreting any single study, regardless of how large its sample size might be.

An important limitation of meta-analysis is its susceptibility to biased effect size estimates as a result of such factors as publication bias (such as the "file drawer" effect due to unpublished studies). Although methodologists continue to develop new methods to identify and adjust for publication bias, concerns remain about how well current methods work. For example, Kromrey & Rendina-Gobioff (2006) conclude that current methods to identify publication bias often either fail to control Type I error rates or else lack power. Furthermore, Kraemer et al. (1998) have shown that including underpowered studies in a meta-analysis can create bias, underscoring the importance of designing individual studies with sufficient power. Beyond problems caused by entire studies not being reported, a related problem is selective reporting of results, even in published studies. Chan et al. (2004) find clear evidence of selective reporting within studies in a large literature review they conducted. Chan et al. (2004) recommend that studies should be registered and protocols published online prior to the actual execution of a study. Toward this goal, the member journals of the International Committee of Medical Journal Editors adopted a policy in 2004 requiring registration of all clinical trials in a public

trials registry as a condition of consideration for publication.

Multisite studies offer an alternative to meta-analysis and at least in principle are less prone to the "file drawer" effect. Kelley & Rausch (2006, p. 375) point out that "The idea of such multisite studies is to spread the burden but reap the benefits of estimates that are accurate and/or statistically significant." However, Kraemer & Robinson (2005, p. 528) point out that it is important to "prevent premature multicenter RCTs [randomized clinical trials] that may waste limited funding, investigator time and resources, and burden participants for little yield." They discuss various aspects of sample size planning in multicenter studies and provide a model of the respective roles of multicenter studies and individual studies in creating a cumulative science.

SUMMARY AND CONCLUSIONS

Advances continue to be made in methods for sample size planning. Some of these advances reflect analytic contributions for specific statistical methods, whereas others reflect new perspectives on fundamental goals of empirical research. In addition to the developments we have described, other important advances are taking place in the role of pilot studies in sample size planning (e.g., Kraemer et al. 2006), methods for dealing with the fact that the effect size is unknowable prior to conducting a study (e.g., O'Hagan et al. 2005), accounting for uncertainty in sample size due to estimating variance (e.g., Muller & Pasour 1997), adaptive sample size adjustments based on interim analyses (e.g., Jennison & Turnbull 2006, Mehta & Patel 2006), complications in planning subgroup analyses (e.g., Brookes et al. 2004, Lagakos 2006), the impact of noncompliance on power in randomized studies (e.g., Jo 2002), and Bayesian approaches to sample size planning (e.g., Berry 2004, 2006; Inoue et al. 2005; Lee & Zelen 2000). As psychologists consider the importance of effect size measures, it becomes incumbent to recognize the inherent uncertainty in effect size measures observed in small samples. Thus, the AIPE approach to sample size planning should assume an increasing role in psychological research. At the same time, it is important for researchers to appreciate the role of statistical power and AIPE not only for their own individual research but also for the discipline's effort to build a cumulative science.

SUMMARY POINTS

1. Sample size planning is important to enhance cumulative knowledge in the discipline as well as for the individual researcher.
2. Sample size planning can be based on a goal of achieving adequate statistical power, or accurate parameter estimates, or both.
3. Researchers are actively involved in developing methods for sample size planning, especially for complex designs and analyses.
4. Sample sizes necessary to achieve accurate parameter estimates will often be larger than sample sizes necessary to detect even a small effect.
5. Sample sizes necessary to obtain accurate parameter estimates or power to detect small effects may often require resources prohibitive to the individual researcher, thus suggesting the desirability of study registries accompanied by meta-analytic methods.

ACKNOWLEDGMENTS

We thank Chrystyna Kouros for help in the preparation of this review, and Lyle Jones, Helena Kraemer, Russell Lenth, Gitta Lubke, Robert MacCallum, Keith Muller, and Ralph O'Brien for their helpful comments on an earlier version of this review.

LITERATURE CITED

Algina J, Keselman HJ. 2000. Cross-validation sample sizes. *Appl. Psychol. Meas.* 24(2):173–79

Algina J, Moulder BC, Moser BK. 2002. Sample size requirements for accurate estimation of squared semipartial correlation coefficients. *Multivariate Behav. Res.* 37(1):37–57

Algina J, Olejnik S. 2000. Conducting power analyses for ANOVA and ANCOVA in between-subjects designs. *Eval. Health Prof.* 26(3):288–314

Algina J, Olejnik S. 2003. Sample size tables for correlation analysis with applications in partial correlation and multiple regression analysis. *Multivariate Behav. Res.* 38(3):309–23

Am. Educ. Res. Assoc. 2006. *Standards for Reporting on Empirical Social Science Research in AERA Publications*. Washington, DC: Am. Educ. Res. Assoc. **http://www.aera.net/opportunities/?id=1850**

Am. Psychol. Assoc. 2001. *Publication Manual of the American Psychological Association*. Washington, DC: Am. Psychol. Assoc. 5th ed.

Barthel FMS, Babiker A, Royston P, Parmar MKB. 2006. Evaluation of sample size and power for multi-arm survival trials allowing for nonuniform accrual, nonproportional hazards, loss to follow-up and cross-over. *Stat. Med.* 25:2521–42

Bates DM, Watts DG. 1988. *Nonlinear Regression Analysis and Its Applications*. New York: Wiley

Beal SL. 1989. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* 45(3):969–77

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57(1):289–300

Bernardo MVP, Harrington DP. 2001. Sample size calculations for the two-sample problem using the multiplicative intensity model. *Stat. Med.* 20:557–79

Berry DA. 2004. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat. Sci.* 19(1):175–87

Berry DA. 2006. Bayesian clinical trials. *Nature* 5:27–36

Bezeau S, Graves R. 2001. Statistical power and effect sizes of clinical neuropsychology research. *J. Clin. Exp. Neuropsychol.* 23:399–406

Biesanz JC, Deeb-Sossa N, Papadakis AA, Bollen KA, Curran PJ. 2004. The role of coding time in estimating and interpreting growth curve models. *Psychol. Methods* 9(1):30–52

Bollen KA, Curran PJ. 2006. *Latent Curve Models: A Structural Equation Perspective*. Hoboken, NJ: Wiley

Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. 2004. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J. Clin. Epidemiol.* 57:229–36

Browne MW, Cudeck R. 1993. Alternative ways of assessing model fit. In *Testing Structural Equation Models*, ed. K Bollen, S Long, pp. 136–62. Newbury Park, NJ: Sage

Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. 2004. Sample size calculator for cluster randomized trials. *Comput. Biol. Med.* 34:113–25

Cashen LH, Geiger SW. 2004. Statistical power and the testing of null hypotheses: a review of contemporary management research and recommendations for future studies. *Organ. Res. Methods* 7(2):151–67

Chan A. 2002. Power and sample size determination for noninferiority trials using an exact method. *J. Biopharm. Stat.* 12(4):457–69

Chan A, Altman DG. 2005. Epidemiology and reporting of randomized trials published in PubMed journals. *Lancet* 365:1159–62

Chan A, Hróbjartsson A, Haahr MT, Gotzsche PC, Altman DG. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols for published articles. *JAMA* 2291(20):2457–65

Coffey CS, Muller KE. 2003. Properties of internal pilots with the univariate approach to repeated measures. *Statist. Med.* 22:2469–85

Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65(3):145–53

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

Cohen J. 1994. The earth is round ($p < .05$). *Am. Psychol.* 49(12):997–1003

Cohn LD, Becker BJ. 2003. How meta-analysis increases statistical power. *Psychol. Methods* 8(3):243–53

Diggle PJ, Heagerty P, Liang KY, Zeger SL. 2002. *Analysis of Longitudinal Data*. Oxford: Oxford Univ. Press. 2nd ed.

Dolan C, Van Der Sluis S, Grasman R. 2005. A note on normal theory power calculation in SEM with data missing completely at random. *Struct. Equ. Model.* 12(2):245–62

Federov V, Jones B. 2005. The design of multicentre trials. *Stat. Methods Med. Res.* 14:205–48

Festinger L, Carlsmith JM. 1959. Cognitive consequences of forced compliance. *J. Abnorm. Soc. Psychol.* 58(2):203–10

Glueck DH, Muller KE. 2003. Adjusting power for a baseline covariate in linear models. *Stat. Med.* 22:2535–51

Goodman SN, Berlin JA. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann. Intern. Med.* 121:200–6

Green SB. 1991. How many subjects does it take to do a regression analysis? *Multivariate Behav. Res.* 26:499–510

Greenwald AG. 1975. Consequences of prejudice against the null hypothesis. *Psychol. Bull.* 82:1–20

Hancock GR. 2006. Power analysis in covariance structure modeling. In *Structural Equation Modeling: A Second Course*, ed. GR Hancock, RO Mueller, pp. 69–115. Greenwich, CT: Inf. Age Publ.

Hancock GR, Freeman MJ. 2001. Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educ. Psychol. Meas.* 61(5):741–58

Harlow LL, Mulaik SA, Steiger JH. 1997. *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum

Hedeker D, Gibbons RD, Waternaux C. 1999. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *J. Educ. Behav. Stat.* 24(1):70–93

Hedges LV, Pigott TD. 2001. The power of statistical tests in meta-analysis. *Psychol. Methods* 6(3):203–17

Hedges LV, Pigott TD. 2004. The power of statistical tests for moderators in meta-analysis. *Psychol. Methods* 9(4):426–45

Hernandez AV, Steyerberg EW, Habbema DF. 2004. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J. Clin. Epidemiol.* 57:454–60

Hoenig JM, Heisey DM. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Stat.* 55(1):19–24

Hogarty KY, Hines CV, Kromrey JD, Ferron JM, Mumford KR. 2005. The quality of factor solutions in explanatory factor analysis: the influence of sample size, communality, and overdetermination. *Educ. Psychol. Meas.* 65(2):202–26

Hsieh FY, Bloch DA, Larsen MD. 1998. A simple method of sample size calculation for linear and logistic regression. *Stat. Med.* 17:1623–34

Hsu JC. 1989. Sample size computation for designing multiple comparison experiments. *Comput. Stat. Data Anal.* 7:79–91

Hsu JC. 1996. *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall

Hunter JE, Schmidt FL. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage. 2nd ed.

Ioannidis JPA. 2005. Why most published research findings are false. *PLoS Med.* 2(8):696–701

Inoue LY, Berry DA, Parmigiani G. 2005. Relationship between Bayesian and frequentist sample size determination. *Am. Stat.* 59(1):79–87

Jennison C, Turnbull BW. 2006. Adaptive and nonadaptive group sequential tests. *Biometrika* 93(1):1–21

Jiroutek MR, Muller KE, Kupper LL, Stewart PW. 2003. A new method for choosing sample size for confidence interval-based inferences. *Biometrics* 59:580–90

Jo B. 2002. Statistical power in randomized intervention studies with noncompliance. *Psychol. Methods* 7(2):178–93

Jones LV, Tukey JW. 2000. A sensible formulation of the significance test. *Psychol. Methods* 5(4):411–14

Julious SA. 2004. Sample sizes for clinical trials with normal data. *Stat. Med.* 23:1921–86

Jung S, Ahn C. 2003. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Stat. Med.* 22:1305–15

Jung S, Ahn C. 2005. Sample size for a two-group comparison of repeated binary measurements using GEE. *Stat. Med.* 24:2583–96

Kelley K. 2007. Sample size planning for the squared multiple correlation coefficient: accuracy in parameter estimation via narrow confidence intervals. *Behav. Res. Methods*. In press

Kelley K, Maxwell SE. 2003. Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychol. Methods* 8(3):305–21

Kelley K, Maxwell SE. 2008. Sample size planning with applications to multiple regression: power and accuracy for omnibus and targeted effects. In *The Sage Handbook of Social Research Methods*, ed. P Alasuutari, L Bickman, J Brannen, pp. 166–92. London: Sage

Kelley K, Maxwell SE, Rausch JR. 2003. Obtaining power or obtaining precision: delineating methods of sample-size planning. *Eval. Health Prof.* 26(3):258–87

Kelley K, Rausch JR. 2006. Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. *Psychol. Methods* 11(4):363–85

Kim KH. 2005. The relation among fit indexes, power, and sample size in structural equation modeling. *Struct. Equ. Model.* 12(3):368–90

Kraemer HC, Gardner C, Brooks JO, Yesavage JA. 1998. The advantages of excluding underpowered studies in meta-analysis: inclusionist vs exclusionist viewpoints. *Psychol. Methods* 3:23–31

Kraemer HC, Mints J, Noda A, Tinklenberg J, Yesavage JA. 2006. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch. Gen. Psychiatry* 63:484–89

Kraemer HC, Robinson TN. 2005. Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? *Contemp. Clin. Trials* 26:518–29

Kromrey JD, Rendina-Gobioff G. 2006. On knowing what we do not know. *Educ. Psychol. Meas.* 66(3):357–73

Lagakos SW. 2006. The challenge of subgroup analyses—reporting without distorting. *N. Engl. J. Med.* 354:1667–70

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med.* 327:248–54

Lee SJ, Zelen M. 2000. Clinical trials and sample size considerations: another perspective. Rejoinder. *Stat. Sci.* 15(2):108–10

Lei PW, Dunbar SB. 2004. Effects of score discreteness and estimating alternative model parameters on power estimation methods in structural equation modeling. *Struct. Equ. Model.* 11(1):20–44

Leon AC. 2004. Sample size requirements for comparisons of two groups on repeated observations of a binary outcome. *Eval. Health Prof.* 27(1):34–44

Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22

Lipsey MW. 1990. *Design Sensitivity: Statistical Power for Experimental Research*. Thousand Oaks, CA: Sage

Liu J, Hsueh H, Hsieh E, Chen JJ. 2002. Tests for equivalence or noninferiority for paired binary data. *Stat. Med.* 21:231–45

Lubke G, Neale MC. 2006. Distinguishing between latent classes and continuous factors: resolution by maximum likelihood? *Multivariate Behav. Res.* 41:499–532

Lyles RH, Lin H, Williamson JM. 2007. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Stat. Med.* 26:1632–48

MacCallum RC, Browne MW, Cai L. 2006. Testing differences between nested covariance structure models: power analysis and null hypotheses. *Psychol. Methods* 11(1):19–35

MacCallum RC, Browne MW, Sugawara HM. 1996. Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* 1(2):130–49

MacCallum RC, Hong S. 1997. Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behav. Res.* 32(2):193–210

MacCallum RC, Widaman KF, Preacher KJ, Hong S. 2001. Sample size in factor analysis: the role of model error. *Multivariate Behav. Res.* 36(4):611–37

MacCallum RC, Widaman KF, Zhang S, Hong S. 1999. Sample size in factor analysis. *Psychol. Methods* 4(1):84–99

Maggard MA, O'Connell JB, Liu JH, Etzioni DA, Ko CK. 2003. Sample size calculations in surgery: Are they done correctly? *Surgery* 134(2):275–79

Maki E. 2006. Power and sample size considerations in clinical trials with competing risk endpoints. *Pharm. Stat.* 5:159–71

Maxwell SE. 2000. Sample size and multiple regression analysis. *Psychol. Methods* 5(4):434–58

Maxwell SE. 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol. Methods* 9(2):147–63

McArdle JJ. 1988. Dynamic but structural equation modeling of repeated measures data. In *The Handbook of Multivariate Experimental Psychology*, Vol. 2, ed. JR Nesselroade, RB Cattell, pp. 561–614. New York: Plenum

McArdle JJ, Epstein D. 1987. Latent growth curves within developmental structural equation models. *Child. Dev.* 58:110–33

McCartney K, Rosenthal R. 2000. Effect size, practical importance, and social policy for children. *Child. Dev.* 71(1):173–80

McClelland GH. 2000. Increasing statistical power without increasing sample size. *Am. Psychol.* 55(8):963–64

McLachlan GJ, Peel D. 2000. *Finite Mixture Models*. New York: Wiley

Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46(4):806–34

Mehta CR, Patel NR. 2006. Adaptive, group sequential and decision theoretic approaches to sample size determination. *Stat. Med.* 25:3250–69

Meredith W, Tisak J. 1984. *"Tuckerizing" curves*. Presented at Annu. Meet. Psychometric Soc., Santa Barbara, Calif.

Meredith W, Tisak J. 1990. Latent curve analysis. *Psychometrika* 55:107–22

Miller RG. 1981. *Simultaneous Statistical Inference*. New York: Springer-Verlag. 2nd ed.

M' Lan CE, Joseph L, Wolfson DB. 2006. Bayesian sample size determination for case-control studies. *J. Am. Stat. Assoc.* 101(474):760–72

Moerbeek M. 2006. Power and money in cluster randomized trials: When is it worth measuring a covariate? *Stat. Med.* 25:2607–17

Mooijaart A. 2003. Estimating the statistical power in small samples by empirical distributions. In *New Development in Psychometrics*, ed. H Yanai, A Okada, K Shigemasu, Y Kano, JJ Meulman, pp. 149–56. Tokyo: Springer-Verlag

Muller KE, LaVange LM, Ramey SL, Ramey CT. 1992. Power calculations for general linear multivariate models including repeated measures applications. *J. Am. Stat. Assoc.* 87:1209–26

Muller KE, Pasour VB. 1997. Bias in linear model power and sample size due to estimating variance. *Commun. Stat.: Theory Methods* 26:839–51

Muñoz MA, Acuña JD. 1999. Sample size requirements of a mixture analysis method with applications in systematic biology. *J. Theor. Biol.* 196:263–165

Murray DM, van Horn ML, Hawkins JD, Arthur MW. 2006. Analysis strategies for a community trial to reduce adolescent ATOD use: a comparison of random coefficient and ANOVA/ANCOVA models. *Contemp. Clin. Trials* 27:188–206

Muthén BO, Curran PJ. 1997. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychol. Methods* 24(4):371–402

Muthén LK, Muthén BO. 2002. How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.* 9(4):599–620

Nasser F, Wisenbaker J. 2001. Modeling the observation-to-variable ratio necessary for determining the number of factors by the standard error scree procedure using logistic regression. *Educ. Psychol. Meas.* 61(3):387–403

Newson R. 2004. Generalized power calculations for generalized linear models and more. *Stata J.* 4(4):379–401

Nickerson RS. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5(2):241–301

O'Brien RG, Castelloe JM. 2007. Sample size analysis for traditional hypothesis testing: concepts and issues. In *Pharmaceutical Statistics Using SAS: A Practical Guide*, ed. A Dmitrienko, C Chuang-Stein, R D'Agostino, pp. 237–71. Cary, NC: SAS

O'Brien RG, Muller KE. 1993. Unified power analysis for t-tests through multivariate hypotheses. In *Applied Analysis of Variance in Behavioral Science*, ed. LK Edwards, pp. 297–344. New York: Marcel Dekker

O'Hagan A, Steves JW, Campbell MJ. 2005. Assurance in clinical trial design. *Pharm. Stat.* 4:187–201

Pan Z, Kupper LL. 1999. Sample size determination for multiple comparison studies treating confidence interval width as random. *Stat. Med.* 18:1475–88

Prentice DA, Miller DT. 1992. When small effects are impressive. *Psychol. Bull.* 112(1):160–64

Proschan MA. 2005. Two-stage sample size re-estimation based on a nuisance parameter: a review. *J. Biopharm. Stat.* 15(4):559–74

Raudenbush SW. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychol. Methods* 2:173–85

Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage. 2nd ed.

Raudenbush SW, Xiao-Feng L. 2001. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychol. Methods* 6(4):387–401

Rochon J. 1998. Application of GEE procedures for sample size calculations in repeated measures experiments. *Stat. Med.* 17:1643–58

Satorra A, Saris WE. 1985. Power of the likelihood ratio test in covariance structure analysis. *Psychometrika* 50(1):83–90

Schmidt FL. 1996. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods* 1:115–29

Schulgen G, Olschewski M, Krane V, Wanner C, Ruff G, Schumacher M. 2005. Sample size for clinical trials with time-to-event endpoints and competing risks. *Contemp. Clin. Trials* 26:386–96

Senn SJ. 2002. Power is indeed irrelevant in interpreting completed studies. *BMJ* 325:1304

Serlin RC, Lapsley DK. 1985. Rationality in psychological research: the good-enough principle. *Am. Psychol.* 40(1):73–83

Serlin RC, Lapsley DK. 1993. Rational appraisal of psychological research and the good-enough principle. In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, ed. G Keren, C Lewis, pp. 199–228. Hillsdale, NJ: Erlbaum

Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin

Shieh G. 2005. Power and sample size calculations for multivariate linear models with randomized explanatory variables. *Psychometrika* 70(2):347–58

Singer JD, Willett JB. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford Univ. Press

Steiger JH. 1990. Structural model evaluation and modification: an interval estimation approach. *Multivariate Behav. Res.* 25:173–80

Steiger JH, Lind JM. 1980. *Statistically Based Tests for the Number of Common Factors*. Presented at Annu. Meet. Psychometric Soc., Iowa City, IA

Sternberg RJ, Williams WW. 1997. Does the Graduate Record Examination predict meaningful success in the graduate training of psychology? A case study. *Am. Psychol.* 52(6):630–41

Strickland PA, Lu S. 2003. Estimates, power and sample size calculations for two-sample ordinal outcomes under before-after study designs. *Stat. Med.* 22:1807–18

Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. 2007. Evidence-based sample size calculations based upon updated meta-analysis. *Stat. Med.* 26:2479–500

Tang M, Tang N, Chan IS, Chan BP. 2002. Sample size determination for establishing equivalence/noninferiority via ratio of two proportions in matched-pair design. *Biometrics* 58:957–63

Taylor AB, West SG, Aiken LS. 2006. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educ. Psychol. Meas.* 66(2):228–39

Tsonaka R, Rizopoulos D, Lesaffre E. 2006. Power and sample size calculations for discrete bounded outcome scores. *Stat. Med.* 25:4241–52

Tryon WW. 2001. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychol. Methods* 6(4):371–86

Vaeth M, Skovlund E. 2004. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Stat. Med.* 23:1781–92

Velicer WF, Fava JL. 1998. Effects of variable and subject sampling on factor pattern recovery. *Psychol. Methods* 3(2):231–51

Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. 2005. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* 58:475–83

Wang F, Gelfand AE. 2002. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Stat. Sci.* 17:193–208

West SG, Biesanz JC, Pitts SC. 2000. Causal inference and generalization in field settings: experimental and quasi-experimental designs. In *Handbook of Research Methods in Social and Personality Psychology*, ed. HT Reis, CM Judd, pp. 40–84. New York: Cambridge Univ. Press

Wilkinson L, Task Force Statistical Inference. 1999. Statistical methods in psychology journals: guidelines and explanations. *Am. Psychol.* 54(8):594–604

Williams VSL, Jones LV, Tukey JW. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *J. Educ. Behav. Stat.* 24:42–69

Wilson DB, Lipsey MW. 2001. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychol. Methods* 6:413–29

Winkens B, Schouten HJA, van Breukelen GJP, Berger MPF. 2006. Optimal number of repeated measures and group sizes in clinical trials with linearly divergent treatment effects. *Contemp. Clin. Trials* 27:57–69

Yan X, Su X. 2006. Sample size determination for clinical trials in patients with nonlinear disease progression. *J. Biopharm. Stat.* 16:91–105

Yuan K, Hayashi K. 2003. Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *Br. J. Math. Stat. Psychol.* 56(1):93–110

Yuan K, Maxwell SE. 2005. On the post hoc power in testing mean differences. *J. Educ. Behav. Stat.* 30(2):141–67

Zheng J, Frey HC. 2004. Quantification of variability and uncertainty using mixture distributions: evaluation of sample size, mixing weights, and separation between components. *Risk Anal.* 24(3):553–71