Nonequivalent controlled pretest-posttest designs are central to evaluation science, yet no practical and unified approach for estimating power in the two most widely used analytic approaches to these designs exists. This article fills the gap by presenting and comparing useful, unified power formulas for ANCOVA and change-score analyses, indicating the implications of each on sample-size requirements. The authors close with practical recommendations for evaluators. Mathematical details and a simple spreadsheet approach are included in appendices.

STATISTICAL POWER FOR NONEQUIVALENT PRETEST-POSTTEST DESIGNS

The Impact of Change-Score Versus ANCOVA Models

J. MICHAEL OAKES
HENRY A. FELDMAN
New England Research Institutes

Statistical power is an important topic for evaluations aiming to infer cause from (quasi) experimental designs.[1] Studies without sufficient power may lead to incorrect inferences and waste scarce resources and might end up doing more harm than good (Donner 1984). In plain terms, if a study is underpowered, evaluators are prevented from knowing why null hypotheses are sustained (i.e., not rejected). Say, for example, that no statistically significant difference is found between treatment and control groups on some endpoint. This may either mean that (a) the treatment program had no effect or (b) there was insufficient power to detect the relationship that does in fact exist. Although evaluators have expressed interest in statistical power (see Bloom

AUTHORS' NOTE: *This project was supported by R01-HL61573 from the National Heart, Lung and Blood Institute. The comments of several colleagues and friends have improved this article. We thank Andy Anderson, Andre Araujo, George Howard, Ken Kleinman, Gregg McHugo, Paul McGovern, Joe Morrissey, Stuart Pocock, George Reed, Kevin Smith, and an anonymous reviewer. The usual caveats apply.*

EVALUATION REVIEW, Vol. 25 No. 1, February 2001 3-28

1995; Lipsey 1990), more tailored research and pedagogical expositions are warranted.

The controlled pretest-posttest design is one of the more powerful evaluation designs because it can isolate causal factors hypothesized to be at work (Bonate 2000; Rossi, Freeman, and Lipsey 1999). The design may also be the most frequently used in all of social science (Cook and Campbell 1979). Yet, confusion remains on how to estimate the sample size required to ensure sufficient statistical power, especially when control groups are not equivalent. Remarkably, none of the well-known statistical power programs (e.g., PASS 2000, nQuery 4, Power and Precision) include comprehensive dedicated modules to analyze data generated from the design(s).

We speculate that the principal reason for the dearth of "canned modules" is the persistent use of analysis of covariance (ANCOVA) for data collected through controlled pretest-posttest designs. The dominance of ANCOVA seems attributable to the early conclusions of psychometricians, who were dismayed with the change-score approach (Allison 1990), and to the accessibility of Reichardt's (1979) important contribution. However, as the dominance of ANCOVA has diminished (Bonate 2000; Allison 1990), practical approaches to sample size/statistical power calculations are required.

Although other formulas for calculating power and/or sample size have been published (cf. Dawson 1998; Hsieh, Bloch, and Larsen 1998; Lipsitz and Parzen 1995; Bloom 1995; Feldman and McKinlay 1994; Frison and Pocock 1992; Self and Mauritsen 1988; Cohen 1988; Donner 1984), few presentations are accessible to practicing evaluators, fewer still apply to the pretest-posttest design so often employed in evaluation science, and fewer yet appreciate that real-world calculations are often done with limited information. The problem of calculating statistical power for analyses of data collected in pretest-posttest designs remains.

This article aims to fill the gap by assisting evaluators less experienced in statistical methods.[2] We first present the pretest-posttest design and discuss the two major methods for data analysis. We then present and compare large-sample power formulas for ANCOVA and change-score analyses, indicating the implications of each on sample-size requirements. We close with practical recommendations for evaluators trying to calculate power for pretest-posttest designs. The mathematical details and derivations of formulae, including a unified model, are included in an appendix. So, too, is an easy-to-implement spreadsheet approach to calculate power for change-score and ANCOVA analyses.

Before proceeding, it seems important to say something about our aims and nomenclature. The results we present have been synthesized from the disciplines of statistics, biostatistics, psychometrics, sociometrics, and econometrics. We note not only great notational variability within fields but sometimes contradictory notation between fields. In fact, it is a struggle to find common nomenclature between any two authors. Because our goal here is to provide practicing evaluators with useful tools, we have tried to write with a unified, intuitive notation, using descriptive definitions wherever possible. For example, throughout this article, we let subscript small *x*s and *y*s indicate unobservable true-score measures and capital *X*s and *Y*s indicate observed measures. In addition, we have sacrificed some generality and mathematical precision for the sake of understandability and practicality. We trust that disciplinary specialists will map our notation to theirs, and methodologists will abstract and generalize as they deem necessary.

THE PROBLEM

THE DESIGN

Campbell and Stanley (1963) provide a clear presentation of the pretest-posttest design, reminding us that it has been in use since the early-20th century. The design is often of great use to evaluators because it can control for all of the major threats to internal validity, such as maturation, selection, and instrumentation (Boruch 1998; Campbell and Stanley 1963). Not only do subjects serve as their own controls, which yields intraindividual change scores, but such change is assessed between two groups, one of which receives some treatment program. Figure 1 diagrams the design.

We let $O_{t,\ pre}$ represent an observation (i.e., measurement) in the treatment group before the treatment is applied, $O_{t,\ post}$ represent an observation in the treatment group after the treatment is applied, T be the program treatment itself, $O_{c,\ pre}$ an observation in the control/comparison group before the treatment is applied, and $O_{c,\ post}$ an observation in the control/comparison group after the treatment is applied.

Notice the *R*s in the figure: They imply that subjects are randomized to treatment or control arms—ideally after baseline measurements are taken. Control over assignment (i.e., randomization) is critical because it permits evaluators to assume that the expected values of all (measured/unmeasured/umeasurable) variables are equal in the treatment and control groups, save the treatment program.

A related design, more frequently employed but typically less powerful, occurs when subjects are not randomized to program treatment or control arms, that is, when the evaluator cannot control the “when” and “to whom” of

$$\begin{array}{cccc} R & \mathrm{O}_{\mathrm{t,\,pre}} & \mathrm{T} & \mathrm{O}_{\mathrm{t,\,post}} \\ \\ R & \mathrm{O}_{\mathrm{c,\,pre}} & & \mathrm{O}_{\mathrm{c,\,post}} \end{array}$$

Figure 1. Controlled Pretest-Posttest Design

exposure to treatment. This design is traditionally called a quasi-experiment (Reichardt and Mark 1998; Cook and Campbell 1979) and is often represented as Figure 1, except that the *R*s are eliminated and a dotted line is placed between the two rows. The dotted line is to show that treatment and control groups are not randomized by the evaluator and they thus should not be considered statistically equivalent (Pocock and Elbourne 2000; Robins and Freidlander 1995; Heckman and Hotz 1989).

STATISTICAL POWER

Statistical power is the probability that a test of the null hypothesis will correctly yield statistical significance when the null hypothesis is, in fact, false. Even more simply put, statistical power is the chance of finding a difference between two groups when such a difference exists. The point is that, because of random variation, just any difference between treatment and comparison groups may not be sufficient for inferring that such a difference actually exists.

Figure 2 is adopted from Lipsey (1990) and diagrams the issue. Although two measured means may literally be different from one another, the error about the means may be so large that the measured difference cannot be declared statistically different. The difference, denoted $\hat{\Delta}$, may be due to chance alone. The power of a statistical test measures the likelihood that a difference between groups is "detectable." Because β is typically employed to represent the probability of not finding a relationship when one exists (i.e.,

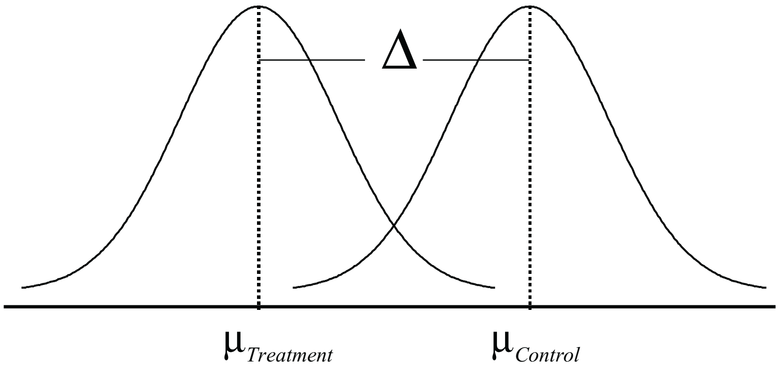


Figure 2. Differences Between Means

Type II error), power is the complement of that error, or (1 – β). If the probability of Type II error is 0.20, then power is 0.80.

Because of the many (often hidden) assumptions involved, power calculations are approximations based on known or estimated facts and typically aim to determine how many subjects are needed to confidently draw conclusions for a particular statistical test. Too few subjects and relationships may be overlooked; too many and scarce resources may be squandered. We intentionally use the word *approximation* because power calculations help us decide whether we need n or $2n$ subjects but not n or $n + 2$. That is, in the practical world, power calculations may tell us if we need 50 or 100 subjects but, because of assumptions, not whether we need 50 or 52 subjects. Five major factors influence statistical power: (a) sample size, (b) variation in the outcome of interest, (c) statistical test to be used, (d) choice of alpha (Type I) error rate, and (e) the actual difference between the groups.

THE PROBLEM

Our experience is that evaluators overlook the influence that an analysis plan (i.e., a particular statistical test) has on power calculations. Indeed, some seem to think that power calculations may be done without reference to both the research design and the specific hypothesis test to be performed. It is a mistake not to tie together the research design, the analysis plan, and power/sample-size calculations for they are inextricably related.

When it comes to the pretest-posttest design addressed here, there is little difficulty when evaluators can randomize subjects to treatment conditions. Randomization at (actually following) baseline allows a simple, independent-sample *t* test to be performed on the two posttest means (although see Bonate 2000 for other options). Power formulas for such a test are readily available in elementary statistics texts and in all the power software we have examined.

For the most part, problems obtain when randomization is not possible. In this case, evaluators often want to account for baseline values of a measure in tests because such values may be different between groups. The issue stems back to debates in the 1960s over whether to analyze data collected from such a design with ANCOVA or change-score models (cf. Maris 1998; Allison 1990; Burr and Nesselroade 1990; Reichardt 1979). Although related, the two models contain subtle differences that may affect (or bias) treatment estimates and always affect power calculations. The problem addressed here obtains because confusion over an analysis plan usually leads to confusion and possibly mistakes in calculating statistical power.

We do not wish to enter the debate over the superiority of analytic approaches. Others are shedding light on the matter, especially for nonrandomized designs (cf. Yanez, Kronmal, and Shemanski 1998; Allison 1990, 1994; Cain, Kronmal, and Kosinski 1992; Rubin 1977). We also acknowledge advances in inferring cause from the data generated through the pretest-posttest design, such as propensity score and selection models (see Winship and Morgan 1999 for an excellent overview), and that subtle assumptions require attention (see Allison 1994; Rosner 1979). But we do not address these more sophisticated or deep assumptions here. Instead, we try to explain the implications of the seemingly most frequently used analytic approaches on sample-size requirements.

ANCOVA MODEL

In the evaluation context addressed here, an ANCOVA aims to estimate a treatment effect on some posttest outcome or impact measure while adjusting for initial pretest scores. This may be done by regressing a posttest score on a pretest score and an indicator variable. The idea is to statistically control (i.e., adjust) for the pretest by means of regression so that one can study the posttest freed of the portion of variance linearly associated with the pretest. The formal model may be written as follows:

$$Y = \alpha + \beta_1 X + \beta_2 T + \varepsilon, \tag{1}$$

where Y is the posttest score on an outcome/impact measure, α is the estimated intercept, X is the pretest score for the same outcome/impact measure, β_1 is the coefficient for the pretest score, T is a (0, 1) indicator variable for treatment or control group, and ε is residual error. The principal null hypothesis is that $\beta_2 = \hat{\Delta} = 0$, which if true suggests that the program being investigated had no effect on the mean of the outcome measure.

The estimated treatment effect for this model may be written as

$$\hat{\Delta} = \hat{\alpha}_t - \hat{\alpha}_c = (\hat{Y}_{t.} - \hat{Y}_{c.}) - \beta(\hat{X}_{t.} - \hat{X}_{c.}), \tag{2}$$

which says that the treatment effect, which is the difference between the modeled intercepts, equals the difference between the predicted posttest means minus the difference between the pretest means multiplied by some coefficient, β.

Among the assumptions in this model, several are especially important to evaluators: (a) There is a linear relationship between X and Y; (b) Y is normally distributed; (c) the residual error variance, ε, is constant for all subjects; (d) the residual errors are independent of one another and the pretest scores; and (e) the pretest score, X, is assumed to be measured without error (Greene 1997). Other, more subtle assumptions such as random sampling and the "stable unit treatment value assumption" (SUTVA) may be found in Winship and Morgan (1999), Rubin (1991), and Holland (1986).

It is the fifth assumption, sometimes called fixed regressors, that is often overlooked and yet causes the most difficulty for evaluators conducting "observational" studies (cf. Yanez, Kronmal, and Shemanski 1998; Chambless and Roeback 1993; Cain, Kronmal, and Kosinski 1992). It is generally known that the incorrectly assuming error-free measurement attenuates the magnitude of the regression coefficient, biasing it toward zero unless errors are independent (Fuller 1987). More simply put, in the presence of pretest measurement error, the coefficient β in Equation 2 is actually multiplied by the psychometric reliability of the pretest, conventionally denoted as ρ_{XX} (Reichardt 1979). Because ρ_{XX} is always less than one, β is reduced and the estimated treatment effect, $\hat{\Delta}$, is altered. In other words, in the presence of measurement error, the slopes of the ANCOVA parallel regression lines become flatter and may decrease or increase their vertical separation, depending on the positions of the treatment and control data. ANCOVA may yield biased results in quasi-experimental designs (cf. Lord 1960).

Importantly, this bias does not obtain if treatments are randomly assigned because that would mean $E(\hat{X}_{t.} - \hat{X}_{c.}) = 0$, which would eliminate any influence of the pretest reliability. Below, we show that the pretest error assumption also has important implications for power calculations.

It is worth pointing out that the ANCOVA model may also be written as

$$(Y - X) = \alpha + \beta_1^* X + \beta_2 T + \varepsilon, \tag{3}$$

where $\beta_1^* = \beta_1 - 1$. This specification, which regresses the difference between the posttest and pretest score on the pretest and treatment indicator variable, yields the same coefficient and standard error for the treatment effect, β_2 (Werts and Linn 1970). Keep this in mind when reading the change-score model below because it implies that one of the key differences between the models is not the dependent variable but the inclusion of a pretest covariate.

CHANGE-SCORE MODEL

Another well-known, although seemingly less frequently employed, approach to analyses of data from pretest-posttest designs is the analysis of change scores, sometimes called gain scores. This model may be written as

$$(Y - X) = \alpha + \beta_2 T + \varepsilon, \tag{4}$$

where Y is the posttest score on an outcome/impact measure, and X is the pretest score on an outcome/impact measure; their difference is the change score. All other symbols are the same as before. The principal null hypothesis remains that $\beta_2 = 0$. Unlike the ANCOVA approach above, this model does not include a pretest covariate on the right-hand side of the equation. Allison (1994) presents this model as the fixed-effect estimator.

This model maintains the same assumptions as the ANCOVA model, except that it does not assume pretest scores are measured without error. In addition, this model assumes that the regression coefficient for posttest scores on pretest scores is unity, that is, $Y = \alpha + X + \beta_2 T + \varepsilon$. This relationship may be more easily seen in terms of the model for the treatment effect in a change-score model:

$$\hat{\Delta} = \hat{\alpha}_t - \hat{\alpha}_c = (\hat{Y}_{t.} - \hat{Y}_{c.}) - (\hat{X}_{t.} - \hat{X}_{c.}). \tag{5}$$

Notice that the only difference between Equations 5 and 2 is the absence of a coefficient for the pretest adjustment. In other words, the coefficient is fixed at unity: $\beta_1 = 1$.

Many of the statistical properties associated with the change-score model spawned a zealous attack by psychometricians and sociometricians (Burr and Nesselroade 1990). Two problems appear to have been central:

1. Reliability: Lord (1967) pointed out that because the denominator in the formula for the reliability of a change score was $2(1 - r_{YX})$, as the correlation between pre- and posttests increased toward unity, the reliability of the change score decreased. This was thought to make change-score values less reliable than their component pre- and posttest values (Burr and Nesselroade 1990).
2. Regression toward the mean: This phenomenon obtains when extreme scores on the pretest move inward toward the mean on the posttest and appears almost universal in the pretest-posttest designs (Allison 1990). The result is that differences are attenuated (Burr and Nesselroade 1990). This problem may be more troublesome in quasi-experimental pretest-posttest designs because there may also be correlation between treatment program and differences.

Since 1975, there has been a tempering of opinion on the use of change-score models (although see Yanez, Kronmal, and Shemanski 1998; Cain, Kronmal, and Kosinski 1992). Allison (1990) shows that (a) the reliability of measures in the change-score model plays no role and introduces no bias because the purported low reliability is due to differencing out the stable component of the measure; and (b) relying on Kenny (1975) and Kenny and Cohen (1979), that regression toward the mean does not pose a problem when evaluators are comparing two stable groups—as is the case with the design discussed here.

Allison (1990) goes on to say that the change-score method is superior to ANCOVA whenever T is subsequent to the pretest and uncorrelated with the "transient" component of the measure. By transient, he means that there is no causal interaction between group and treatment. Or, in still other words, that there is no systematic differential change (i.e., increasing growth) in the treatment group compared to the control. A key aspect of his statement is that the "extra" assumption about the implicit value of the pretest coefficient (i.e., $\beta = 1$) is largely unimportant.

The implicit unity coefficient assumption means that for every one-unit increase in a pretest, we should expect a one-unit increase in a posttest. Statisticians and psychometricians have argued that, if not supported by data,[3] this assumption reduces the usefulness of the change-score model by inflating the variance of the treatment estimate (Reichardt 1979; Feldt 1957). However, Allison (1990) persuasively argues that (a) it is a reasonable assumption to make in (quasi) experimental designs, and (b) even if wrong it (i) probably has little substantive impact and (ii) is a reasonable assumption when causal relationships are complex. Readers familiar with this literature will recognize that Allison's concept is mathematically expressed in Winship and Mare's (1992) Equation 6 as the differential treatment effect, $(\bar{\delta}_{i\varepsilon T} - \bar{\delta}_{i\varepsilon C})$.

POWER FOR THE CHANGE-SCORE MODEL

Contrary to many texts, we think evaluators often begin thinking about power by considering how many subjects they can afford to observe, the conventional desired power for a given test (e.g., 80%), and the desired alpha error (e.g., 5%). Accordingly, instead of calculating power, (1 – β), or sample size, *n*, we often calculate the minimum detectable difference, Δ, between groups for a given test. We thus admit to using the term *power calculations* a bit loosely.

Because it seems easier to understand, we begin with the formula for change-score analysis. The actual minimal detectable effect for a two-group, pretest-posttest controlled design analyzed with a change-score model is

$$\Delta = \sqrt{\frac{4\sigma_Y^2\left(Z_{\alpha/2} + Z_\beta\right)^2\left(1 - R_{Yx}^2\right)}{n}}, \tag{6}$$

where Δ is the minimum detectable difference between groups, σ_Y^2 is the total variance of posttest values, $Z_{\alpha/2}$ is the normal deviate at a selected Type I error rate, Z_β is the normal deviate for a given Type II error rate, *n* is the per-group sample size, and R_{Yx}^2 is the proportion of variance in the observed control-group posttest score explained by the true score of the control-group pretest.

The formula is actually quite simple to use. Because Type I error rates are conventionally 5%, $Z_{\alpha/2}$ is often 1.96. Because Type II error rates are conventionally 20%, power is 80%, and thus Z_β is often 0.84. The squared product of these two terms is therefore 7.84. The variance of the observed posttest score, σ_Y^2, may be estimated from data, previous studies, or as a last resort as 25% of an endpoint's range. Per-group sample size, *n*, is usually determined by budget constraints and may vary to any degree an evaluator wishes. The only challenging term is R_{Yx}^2.

The explanation of R_{Yx}^2 requires some digression. Classical measurement theory (see Nunnally and Bernstein 1994 or Bohrnstedt 1983) is based on the equation

$$x = \tau + \varepsilon, \tag{7}$$

or what is the same in our notation,

$$X = x + \varepsilon. \tag{8}$$

Equation 8 says that an observable response, X, for a given measure is the sum of its unobservable true score, x, and measurement error, ε. We believe that all observed variables are measured with error. Psychometricians estimate this error with reliability statistics, which survey researchers often call intraclass correlations (ICCs). We denote reliability of a measure as R^2_{Xx}, which may be written as

$$R^2_{Xx} = \frac{\sigma^2_{true}}{\sigma^2_{true} + \sigma^2_{error}} = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_\varepsilon}. \tag{9}$$

Notice that the reliability of a measure is a fraction of its total variance and has a range [0,1]. Reliabilities, or ICCs, are routinely published in psychometric studies of endpoints, especially in public health/behavioral medicine studies that rely so often on multi-item scales (cf. McDowell and Newell 1996). Reliabilities above 0.70 are considered acceptable for a new sociometric scale, and 0.90 is considered acceptable for a stable (i.e., state) scale physiologic scale. For new or single-item measures, the reliability R^2_{Xx} may be approximated by R_{XX}, the test-retest Pearson correlation coefficient between pretest and posttest scores. More formally, $R^2_{Xx} \approx R_{XX}$ (Streiner and Norman 1999, p. 115). Because $R^2_{Xx} < R_{XX}$, evaluators may be inclined to subjectively deflate the correlation to better approximate the true reliability value. However, because Bohrstedt (1983) states that observed test-retest reliabilities usually underestimate true reliabilities, deflation may not be necessary. In any case, our experience suggests that the correlation between pretest and posttest measures typically ranges from 0.30 to 0.50 for stable trait assessments over reasonable study periods, making the estimate of R^2_{Xx} fall on the interval [0.09-0.25]. We encourage evaluators to focus more attention on this important area of research.

Not to be confused with R^2_{Xx}, the term R^2_{Yx} may be thought of as a reliability of a true pretest score to an observed posttest score. It is the fraction of the variance of observed posttest score attributable to the true pretest score in the absence of change.

Where can evaluators find a value for R^2_{Yx} ? The term R^2_{Yx} is the quotient of the proportion of variance of the observed posttest explained by the observed pretest to the reliability of the pretest score. Using our notation, this becomes

$$R^2_{Yx} = \frac{R^2_{YX}}{R^2_{Xx}}. \tag{10}$$

Equation 10 means that evaluators can estimate R^2_{Yx} by dividing the proportion of variance explained in the observed posttest by the observed pretest by the reliability of the pretest measure. If a measure is perfectly reliable, then $R^2_{Xx} = 1.0$ and $R^2_{Yx} = R^2_{YX}$. Note, however, that the value of the quotient should not exceed unity: $R^2_{Xx} \geq R^2_{YX}$.

The upshot is that Equation 10 permits evaluators to substitute estimable values into Equation 5 and get a practical formula for the minimal detectable effect for a change-score analysis:

$$\Delta = \sqrt{\frac{4\sigma^2_Y (Z_{\alpha/2} + Z_\beta)^2 (1 - \frac{R^2_{YX}}{R^2_{Xx}})}{n}}. \tag{11}$$

Equation (11) is useful when trying to estimate the minimum detectable effect for a study using change-score analysis. Evaluators need only determine (or intentionally vary for purposes of sensitivity analysis) σ^2_Y, R^2_{Xx}, and R^2_{YX} to conduct a sensitivity analyses for Δ. Sensitivity analyses are critical because they reveal the implications of each term on the minimum detectable difference in a study.

POWER FOR ANCOVA MODEL

The formula approximating Δ for ANCOVA analysis is

$$\Delta = \sqrt{\frac{2\sigma^2_\varepsilon (Z_{\alpha/2} + Z_\beta)^2}{n(1 - R^2_{xG})}}. \tag{12}$$

Most of the terms have already been described above. What is new is the term R^2_{xG} in the denominator and the term σ^2_ε that replaced the σ^2_Y in the numerator.

The term R^2_{xG} symbolizes the proportion of variance in the true pretest, *x*, explained by group membership, *G*. For purposes here, subjects may be in either the treatment or control groups. R^2_{xG} is thus related to selection bias (cf. Berk 1983; Berk and Ray 1982). This term is related to Reichardt's (1979) "D," found in his equation for the precision of the treatment effect estimate. In an experimental (i.e., randomized) design, this value approaches zero, yielding no effect on the expression for Δ. Randomization is in fact what Cohen (1988) assumes in his f′ effect size (ES) for ANCOVA analyses. Note that in studies with extreme selection bias, R^2_{xG} approaches 1 and pushes Δ toward infinity. Power becomes meaningless in studies infected with extreme selection bias.

We typically do not have an estimate of R^2_{xG} in the design (i.e., proposal) phase of a study. Obviously, the objective is to have $R^2_{xG} \rightarrow 0$, which is accomplished through randomization—although it can be shown that as $n \rightarrow 0$, $R^2_{xG} \neq 0$. Evaluators relying on a quasi-experimental design may conduct sensitivity analyses for Δ by letting R^2_{xG} vary over some reasonable range (cf. Rosenbaum 1995). We tentatively speculate that for a study with a carefully selected comparison group, R^2_{xG} may vary over the interval [0.01-0.10]. Post hoc publication of this value by several evaluators would improve our science.

The other new term, σ^2_ε, symbolizes the variance of measurement error and seems more difficult to grasp. σ^2_ε is not the total variance of the primary endpoint: $\sigma^2_\varepsilon \neq \sigma^2_Y$. Rather, σ^2_ε is the unobservable error of the posttest measure. Those familiar with classical psychometric test theory will recognize this term, but even they might find it difficult to estimate during the design phase of a study. A solution is possible, however. Following Bohrnstedt (1983, Equation 3.12), we know that

$$\sigma^2_\varepsilon = \sigma^2_Y(1 - R^2_{Yx}), \tag{13}$$

where σ^2_Y is the total variance of observed posttest and R^2_{Yx} is defined in Equation 10 above.

The upshot of Equation 13 is that it permits evaluators to substitute known quantities into Equation 12 and, as before, use Equation 10 to get a useful formula. For practical purposes, then, we may thus write the power formula for ANCOVA as

$$\Delta = \sqrt{\frac{2\sigma^2_y(Z_{\alpha/2} + Z_\beta)^2(1 - \frac{R^2_{YX}}{R^2_{Xx}})}{n(1 - R^2_{xG})}}, \tag{14}$$

where all the terms are now recognizable, estimable, or reasonably varied over some range.

Recall from Equation 10 that the ANCOVA models assume that the pretest regressor, X, is measured with perfect precision, that is, $R^2_{Xx} = 1.0$. This makes $R^2_{Yx} = R^2_{YX}$ and $R^2_{xG} = R^2_{XG}$. Which means that the formula for ANCOVA is effectively

$$\Delta = \sqrt{\frac{2\sigma^2_Y(Z_{\alpha/2} + Z_\beta)^2(1 - R^2_{YX})}{n(1 - R^2_{XG})}}, \tag{15}$$

which may be familiar to those working in this area.

CHANGE-SCORE VERSUS ANCOVA

The preceding discussion raises several important questions: What are the differences between the power formulae of equations for the change-score model (Equation 11) and the ANCOVA model (Equation 15), and what do such differences mean?

The two equations differ in three key areas: First, the numerator in the change-score model is multiplied by four instead of two. The reason for this is that the ANCOVA model of Equation 15 assumes that regressors are measured without error, which eliminates a variance component. Second, Equation 15 includes the term R^2_{XG} in the denominator and Equation 11 does not. This means that power calculations for ANCOVA analyses are affected by selection bias issues (i.e., difference in groups at baseline), whereas calculations for change-score analyses are not. Pretest group differences are eliminated through the differencing of change-score analyses. Finally, the change-score model incorporates the reliability of the pretest, R^2_{Xx}, whereas the ANCOVA model assumes perfect reliability.

What are the implications of these differences? If we assume a perfectly reliable pretest measure ($R^2_{Xx} = 1$) and a randomized experiment ($R^2_{XG} = 0$), then it turns out that the proportion of variance in the posttests accounted for by the pretest, R^2_{XY}, plays no role in the formula for change-score analyses. Equation 11 thus yields detectable differences or ESs that are 29.3% larger than the formula for ANCOVA, Equation 15, all else equal. This means that in an experimental situation with perfect measurement, ANCOVA is about 30% more precise than change-score analysis. This is a good thing and widely recognized (see Reichardt 1979). But if change-score analysis is the preferred analytic method—and according to Allison (1990), it may well be in quasi-experiments—this also means that studies powered for ANCOVA that end up using use change-score analyses will be about 30% less precise, all else equal.

As the correlation between pretest measures and group assignment, R^2_{xG}, increases from 0, the magnitude of the minimum detectable effect, Δ, for the ANCOVA model increases and the power differences between change-score and ANCOVA decrease, still assuming perfect measures ($R^2_{Xx} = 1$). It is not difficult to show that when $R^2_{xG} = 0.5$, the two equations yield identical output: As $R^2_{xG} \to 0.5$, the difference in Δ between ANCOVA and change-score models becomes negligible. When $R^2_{xG} > 0.5$, Δ is greater in ANCOVA than in change-score analyses. All of this should be intuitively obvious: Selection bias increases the minimum detectable difference between two groups. Because a smaller Δ is better, the presence of selection bias in quasi-experimental studies increases the number of required subjects for a given Δ, which increases the cost. Randomization leads to the least expensive inferences, all else equal. Thus, not only may selection bias lead to improper inferences in the traditional sense (cf. Berk 1983), evaluators employing ANCOVA to analyze data from quasi-experiments may estimate biased results from the intrinsic properties of the model itself. And because the required sample size is increased due to the presence of selection bias, there is a cost premium to pay to the improper analysis of less than perfect data.

Although it is the real-world case, things become a great deal more complicated if we permit measurement error into the models, that is, letting $R^2_{Xx} < 1$, because comparative answers depend on the magnitude of R^2_{YX} and the relations between unmeasurable true scores. Furthermore, if data do not conform to the change-score model's implicit assumption and the regression coefficient between pretest and posttest score departs from unity, then the change-score model is not applicable and should not be used. These relationships are for another article, and we will say nothing more about this here.

EFFECT SIZE

We note one more aspect of power calculations before concluding. The minimal detectable difference, Δ, is based on the metric scale (e.g., pounds, dollars, points) of the primary endpoint, as reflected in σ_Y. Because scales may differ across studies, Δ is difficult to compare across studies. A standardized (i.e., unitless) measure is needed.

As described by Lipsey (1990) and Cohen (1988), it is often useful to standardize the scale of Δ by dividing it by the standard deviation of the pooled prestest, σ_X. This parameter is called the ES (effect size) and may be used to compare power, or detectable effects, across a range of studies with different measurement units. Because outcome variables for a novel study may be arbitrarily scaled, ES becomes an important parameter in discussions of statistical power. ES may be written as

$$\textit{Effect Size} = ES = \frac{\Delta}{\sigma_X}. \tag{16}$$

Researchers believe that for the behavioral sciences, ESs valued at 0.20 may be categorized as small, 0.50 medium, and 0.80 large (Lipsey 1990; Cohen 1988). The smaller the desired ES, the more powerful a study needs to be. In other words, more powerful studies have a better chance at detecting smaller differences between treatment and control groups. The smaller ES we can detect the better.

DISCUSSION AND CONCLUSION

Nonequivalent control-group pretest-posttest designs are central to evaluation science, yet no practical and unified approach for estimating power in the two most widely used analytic approaches exists. This article filled the gap by presenting and comparing useful large-sample power formulas for ANCOVA and change-score analyses, indicating the implications of each on sample-size requirements. We discussed the models, their assumptions, and the impacts of such assumptions on power. Our take-home message is that statistical power calculations need to be related to the research design and the statistical test(s) of interest.

The principal finding is that for a randomized experiment, ANCOVA yields unbiased treatment estimates and typically has superior power to change-score methods, all else equal. However, in the absence of randomization, when baseline differences between groups exist, we follow Allison (1990) and show that change-score models yield less biased estimates (if biased at all). Then, bias aside, we went on to show that the common assumption that ANCOVA models are more powerful rests on the untenable assumption that pretests are measured without error. In the presence of measurement error, change-score models may be equally or even more powerful.

Although the debate over analytic methods will no doubt continue, these findings should help practicing evaluators determine which analytic model is appropriate and how to calculate power for them.

Before closing, we offer a few general and subjective comments about statistical power.

1. Statistical power is critically important to evaluation science. Underpowered studies may lead to wrong and perhaps damaging inferences. The added expense of adding more subjects to a study will pale in comparison to the loss of reputation due to incorrect inferences due to an underpowered study.
2. More subjects are generally better, but randomization, if possible, is critical to drawing proper inferences. Differences in groups at baseline may bias estimates, especially where ANCOVA is employed.
3. Errors in the measurement of the outcome variables weaken and may negate the inferences. It is critically important to use reliable measures and precisely measure variables consistently across subjects and over time, especially when using ANCOVA models.
4. Attrition will occur in all longitudinal studies and must be accounted for. All sample-size formulae and requirements herein are for postattrition levels. If you expect a 25% attrition rate, then you need to inflate any sample size by 1.25, at least. Expected attrition rates should be defended.

5. Differential nonrandom attrition may occur because of the difference in personal attention evaluators pay to treatment and control groups. Differential attrition is not random and not ignorable (cf. Foster and Bickman 1996). As it can bias estimates, it is critical to minimize this effect.
6. Stronger interventions should have larger impacts and thus permit larger thresholds of detectable effects and smaller sample sizes. Nothing is as good as a strong, focused intervention.
7. The more statistical tests you perform, the more likely you will find a difference when none really exists. This phenomenon is called alpha decay and results from "mining" the data for any statistically significant difference (cf. Pocock, Geller, and Tsiatis 1987). It is critically important to plan your primary tests in advance and stick to them. Secondary analyses, unless planned for, will inflate the error rates and should be considered exploratory only (Assman et al. 2000).

APPENDIX A

In this appendix, we outline the mathematics behind the two statistical models discussed above, ANCOVA and change-score, as well as their simplest common generalization, a measurement error model with one dichotomous independent variable and one continuous covariate (Fuller 1987). We specify the assumptions underlying each model, provide estimation and standard-error formulas for the intervention effect in each case, and show how apparent discrepancies between the ANCOVA and change-score models follow from differing assumptions.

Sample statistics. N subjects are recruited for an evaluation experiment. The outcome variable X is measured at baseline, after which n_1 subjects are randomly selected for intervention and n_0 for controls. After intervention, a follow-up measurement Y of the outcome variable is made on each subject. Means taken over the entire data are denoted in the usual way by $\overline{X}$, $\overline{Y}$, $\overline{X^2}$, $\overline{Y^2}$, and $\overline{XY}$. Means within the intervention or control group are denoted by $\overline{X}_G$ and $\overline{Y}_G$, where $G = 1$ for intervention subjects and $G = 0$ for controls.

Measurement error model. The observed baseline measurement X is assumed to be the sum of an unobserved ("true") value x and random Gaussian measurement error:

$$X = x + \mathcal{N}(0, \sigma_\varepsilon^2). \tag{A1}$$

The unobserved values x will be treated as random variables and their variance denoted σ_x^2.

The follow-up outcome measurement is assumed to be the sum of three terms: a constant depending on treatment group, a regression term depending on the baseline value ("true," not observed), and a Gaussian deviate due to measurement error at follow-up. Thus,

$$Y = \Delta_G + \beta x + \mathcal{N}(0, \sigma_\varepsilon^2). \tag{A2}$$

The effect of intervention is

$$\Delta = \Delta_1 - \Delta_0. \tag{A3}$$

The total variance of the outcome measurement is

$$\sigma_Y^2 = \beta^2 \sigma_x^2 + \sigma_\varepsilon^2. \tag{A4}$$

Special case: ANCOVA model. If one assumes that the baseline value is measured without error, then the variability of the “true” values x is the only source of random variance at baseline:

$$X = x. \tag{A5}$$

The follow-up measurement can be written in this case with the regression term depending on the observed baseline value X rather than the “true” value:

$$Y = \Delta_G + \beta X + \mathcal{N}(0, \sigma_\varepsilon^2). \tag{A6}$$

Special case: change-score model. Assume that the regression coefficient for dependence of follow-up on baseline is fixed at $\beta = 1$. Then the baseline and follow-up measurements can be written as follows:

$$X = x + \mathcal{N}(0, \sigma_\varepsilon^2). \tag{A7}$$

$$Y = \Delta_G + x + \mathcal{N}(0, \sigma_\varepsilon^2). \tag{A8}$$

For any particular subject, the covariance between baseline and follow-up is due to the simple term x, which can be eliminated by subtraction. The change between baseline and follow-up is thus

$$Y - X = \Delta_G + \mathcal{N}(0, 2\sigma_\varepsilon^2). \tag{A9}$$

Correlation parameters. The two Gaussian error terms in the measurement error model are assumed to be independent of one another and of the unobserved baseline value x. Interdependencies exist, however, among x, the observed values X and Y, and the condition variable G, which can be expressed as follows in terms of the parameters of the measurement error model.

The fraction of variance of X attributable to variation in the underlying “true” value x is an indication of the reliability of the measurement on a given occasion. Commonly termed the reliability coefficient or intraclass correlation (ICC), this quantity is given by

$$R^2_{Xx} = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_\varepsilon}. \tag{A10}$$

In the special case of ANCOVA, where $x = X$, the error variance in Equation A10 vanishes and $R^2_{Xx} = 1$.

The fraction of variance of Y explained by variation in the "true" baseline value x is an indication of the persistence of an individual's characteristics over time:

$$R^2_{Yx} = \frac{\beta^2\sigma^2_x}{\beta^2\sigma^2_x + \sigma^2_\varepsilon}. \tag{A11}$$

In the special case of the change-score model, where β is fixed at 1, Equation A11 reduces to Equation A10 so that $R^2_{Yx} = R^2_{Xx}$.

The fraction of variance of the unobserved baseline value x explained by the condition variable G is an indication of imbalance between the intervention and control subjects with respect to the baseline value:

$$R^2_{xG} = \frac{N\sigma^2_x - n_0\sigma^2_{x(0)} - n_1\sigma^2_{x(1)}}{N\sigma^2_x} = \frac{n_0 n_1(\mu_0 - \mu_1)^2}{N^2\sigma^2_x}, \tag{A12}$$

where μ_G and $\sigma^2_{x(G)}$ denote the mean and variance of unobserved values within treatment groups. In an ideal experiment, random selection of the intervention and control groups would cause this quantity to be zero. In observational studies, where subjects may be steered by unknown confounding forces into either of the two groups to be compared, R^2_{xG} is an indication of selection bias.

The correlation parameters defined in Equation A10 through A12 will enter into the formulas for precision, power, and sample size developed below. In most design exercises, they can be estimated theoretically or approximated from earlier studies. In some cases, however, an investigator may have directly pertinent data, which would necessarily involve the observed values X rather than the unobservable x. The available quantities would be the fraction of variance of Y explained by X, denoted R^2_{YX}, and the fraction of variance of X explained by G, denoted R^2_{XG}. The quantities in Equations A11 and A12 are readily estimated from the corresponding data-based parameters:

$$R^2_{Yx} = \frac{R^2_{YX}}{R^2_{Xx}} \tag{A13}$$

$$R^2_{xG} = \frac{R^2_{XG}}{R^2_{Xx}}. \tag{A14}$$

A data-based estimate of the reliability coefficient R^2_{Xx} would have to be obtained from a study designed for that purpose, for example, short-term replication of the outcome measurement, from which R^2_{Xx} would be estimated as the ratio of within-subject to total variance. In ANCOVA, where one assumes $x = X$ and $R^2_{Xx} = 1$, distinguishing R^2_{YX} from R^2_{Yx} and R^2_{XG} from R^2_{xG} is not necessary.

Parameter estimates. For the measurement error model, an optimal estimate of the regression coefficient is

$$\hat{\beta} = \frac{N\overline{XY} - n_0\overline{X}_0\overline{Y}_0 - n_1\overline{X}_1\overline{Y}_1}{N\overline{X^2} - n_0\overline{X}_0^2 - n_1\overline{X}_1^2 - N\lambda\sigma_\varepsilon^2}, \tag{A15}$$

where λ is the minimum solution of a quadratic equation expressed in determinantal form as

$$0 = \begin{vmatrix} \overline{Y^2} - \lambda\sigma_\varepsilon^2 & \overline{Y} & \overline{Y}_1 & \overline{XY} \\ \overline{Y} & 1 & n_1 \div N & \overline{X} \\ \overline{Y}_1 & n_1 \div N & n_1 \div N & \overline{X}_1 \\ \overline{XY} & \overline{X} & \overline{X}_1 & \overline{X^2} - \lambda\sigma_\varepsilon^2 \end{vmatrix} \tag{A16}$$

(Fuller 1987). In the special case of ANCOVA, where the baseline value is assumed to have no measurement error, the final term in the denominator of Equation A15 vanishes and the expression becomes the standard estimator for parallel-line regression (Zar 1984). For both the general measurement error model and the special case of ANCOVA, the optimal estimate of the intervention effect is

$$\hat{\Delta} = (\overline{Y}_1 - \overline{Y}_0) - \hat{\beta}(\overline{X}_1 - \overline{X}_0). \tag{A17}$$

In the change-score model, $\hat{\beta}$ is not defined because β does not enter the model. The optimal estimator of Δ is

$$\hat{\Delta} = (\overline{Y}_1 - \overline{Y}_0) - (\overline{X}_1 - \overline{X}_0). \tag{A18}$$

Standard error for intervention effect. Complete details concerning variance-covariance estimation for the measurement error model are given by Fuller (1987). The important item for present concerns is the variance of the intervention effect estimate:

$$SE^2(\hat{\Delta}) = \frac{\sigma_Y^2(1-R^2_{Yx})}{1-R^2_{xG}}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\left[1+\frac{1-R^2_{Xx}}{R^2_{Xx}}\left(\frac{R^2_{Yx}}{1-R^2_{Yx}}+\frac{R^2_{xG}}{1-R^2_{xG}}\right)\right]. \tag{A19}$$

This formula applies to the measurement error and ANCOVA models, for which $\hat{\Delta}$ is defined by Equation A17. For ANCOVA, the bracketed term vanishes because $R^2_{Xx} =$ 1:

$$SE^2(\hat{\Delta}) = \frac{\sigma_Y^2(1-R^2_{Yx})}{1-R^2_{xG}}\left(\frac{1}{n_0}+\frac{1}{n_1}\right). \tag{A20}$$

Table A1: Variance of Estimated Intervention Effect (SE^2 (Δ)) in Important Special Cases

	ANCOVA (R^2_{Xx} = 1)	*Change Score* ($R^2_{Xx} = R^2_{Yx}$)
(a) General case	$\frac{\sigma_Y^2(1-R^2_{Yx})}{1-R^2_{xG}}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)$	$2\sigma_Y^2(1-R^2_{Yx})\left(\frac{1}{n_0}+\frac{1}{n_1}\right)$
(b) No confounding (R^2_{xG} = 0)	$\sigma_Y^2(1-R^2_{Yx})\left(\frac{1}{n_o}+\frac{1}{n_1}\right)$	—[a]
(c) Equal samples (n_0 = n_1 = n)	$\frac{2\sigma_Y^2(1-R^2_{Yx})}{n(1-R^2_{xG})}$	$\frac{4\sigma_Y^2(1-R^2_{Yx})}{n}$
(b) and (c)	$\frac{2\sigma_Y^2(1-R^2_{Yx})}{n}$	—[a]

a. Same as above; under change-score model, variance of Δ is not affected by R^2_{xG}.

For the change-score model, $\hat{\Delta}$ is supplied by Equation A18, and the variance is

$$SE^2(\hat{\Delta}) = 2\sigma_Y^2\left(1-R^2_{Yx}\right)\left(\frac{1}{n_0}+\frac{1}{n_1}\right). \tag{A21}$$

Important special cases of Equations A19 through A21 are displayed in Table A1. In particular, when the intervention and control samples are of equal size ($n_0 = n_1 = n$) and randomly selected (R^2_{xG} = 0), the variance is

$$SE^2(\hat{\Delta}) = \frac{2\sigma_Y^2(1-R^2_{Yx})}{n}\times\begin{cases}1 & \text{for ANCOVA;}\\ 1+\frac{1-R^2_{Xx}}{R^2_{Xx}}\frac{R^2_{Yx}}{1-R^2_{Yx}} & \text{for measurement error model;}\\ 2 & \text{for change}-\text{score analysis.}\end{cases} \tag{A22}$$

The smaller apparent variance for ANCOVA is a direct consequence of assuming perfect reliability at baseline. Whenever baseline reliability (R^2_{Xx}) is greater than pre-post correlation (R^2_{Yx})—a typical situation because the pretest measurement is closer in time to the "true" baseline value—the measurement error variance in Equation A22 will be intermediate in magnitude between the two special cases.

*Power, sample size, and detectable effect.*The equation relating least detectable intervention effect to precision, power, and sample size is

$$\delta = SE(\hat{\Delta})(z_{\alpha/2} + z_{\beta}), \tag{A23}$$

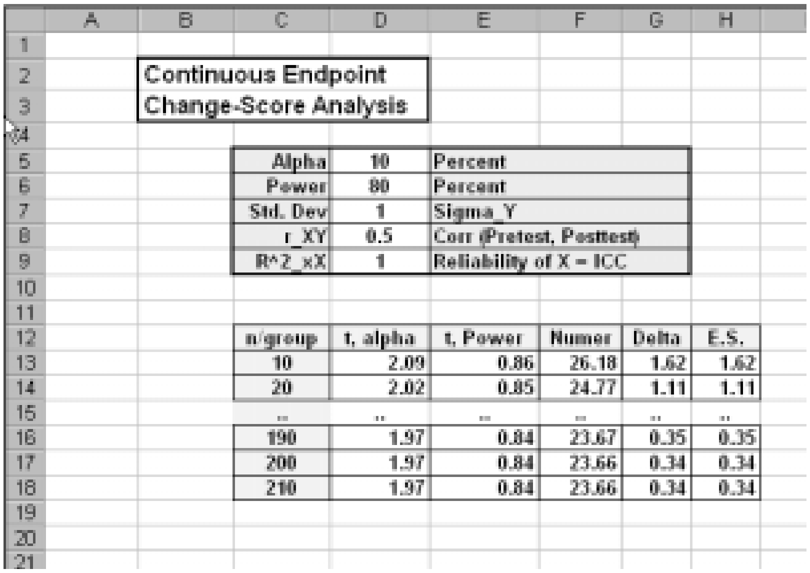
where α and β denote Type I and Type II error rates, respectively; z_p is defined by $p = \int_{-\infty}^{z_p} \mathcal{N}(0,\ 1)$ and power is 100% × (1 – b). In designing a planned evaluation study, any appropriate version of the standard-error formula for $\hat{\Delta}$, as given in Table A1, can be inserted into Equation A23 and solved for whatever power, sample-size, or detectable-effect parameter is required.

APPENDIX B

The formula presented in Equations 11 and 15 should be useful to evaluators conducting power analyses in the design stage of a study. Although many approaches are possible, this section presents one simple way to operationalize the change-score formula in a spreadsheet. Similar spreadsheets may be written for ANCOVA models, as per Equation 15. What follows is written for Microsoft's Excel-97 for Windows software. We use brackets to indicate a cell formula; they are not to be typed in cells.

As shown in Equation 11, there are four key design parameters: α, 1 – β, σ_Y^2, R_{YX}^2, R_{Xx}^2. For practical purposes, these may be translated as (a) Type I error rate, (b) power, (c) standard deviation of the endpoint, and (d) the reliability of the endpoint. Estimates of these parameters are discussed above.

1. The first step is to insert a 2 × 4 matrix in a spreadsheet with labels in the first column and values in the second. We usually also find it useful to add a third column with descriptive comments. For purposes here, let the labels be inserted in cells C5 through C9, with numeric values in D5 through D9.
2. The next step is to type a list of reasonable (i.e., affordable) per-group sample-size values, say 30 to 200 by 10. Label the column "*n* per group," and fill in column C. Begin at C13 and end at C32.
3. In column D, insert a column for the normal deviates of the Type I error rate. Fill cells in this column with the formula [=TINV((($D$5/100)/2),2*C13-1)], where cell D5 contains a percentage (i.e., 10) for the total alpha error.
4. Do likewise in column E for power, filling the cell with [=TINV((((100-$D$6)*2)/100), 2*C13-1)], where cell D6 contains a percentage (i.e., 80) for power.
5. In column F, calculate the numerator of Equation 9. Fill the cells with the formula [=4*((($D$7)^2)*((D13+E13)^2))*(1-($D$8^2/$D$9))].
6. Let column G contain the minimum detectable effect, Δ, for a given sample size *n* per group in C. Fill cells in column G with the formula [=SQRT(F13/C13)].



Continuous Endpoint
Change-Score Analysis

Alpha	10	Percent
Power	80	Percent
Std. Dev	1	Sigma_Y
r_XY	0.5	Corr (Pretest, Posttest)
R^2_xX	1	Reliability of X = ICC

n/group	t, alpha	t, Power	Numer	Delta	E.S.
10	2.09	0.86	26.18	1.62	1.62
20	2.02	0.85	24.77	1.11	1.11
..	..	..	..	..	..
190	1.97	0.84	23.67	0.35	0.35
200	1.97	0.84	23.66	0.34	0.34
210	1.97	0.84	23.66	0.34	0.34

Figure 3: Spreadsheet for Change-Score Analysis

7. Finally, let column H contain the values of the effect size, ES. Fill the cells of this column with [=G13/$D$7], where again D7 is the standard deviation of the endpoint.

Nothing more is needed. Evaluators may now easily calculate statistical power, or more specifically, the minimum detectable difference between treatment and control groups for change-score analyses of data collected in a controlled pretest-posttest design. Figure 3 shows the results of Steps 1 through 7 above.

NOTES

[1] Because it does not directly incorporate Neyman-Pearson hypothesis testing, power is not formally addressed in Bayesian statistical methods, although the concept is somewhat related to highest posterior density intervals. See Barnett 1999 for more information.

[2] We are working on a more rigorous treatment.

3. The assumption can be tested with the data.

REFERENCES

Allison, Paul D. 1990. Change scores as dependent variables in regression analysis. In *Sociological methodology 1990*. Vol. 20, edited by C. Clogg, 93-114. Oxford, UK: Blackwell.

———. 1994. Using panel data to estimate the effects of events. *Sociological Methods and Research* 23:174-99.

Assman, Susan F., Stuart J. Pocock, Laura E. Enos, and Linda E. Kasten. 2000. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* 355:1064-69.

Barnett, Vic. 1999. *Comparative statistical inference*. New York: John Wiley.

Berk, Richard A. 1983. An introduction to sample selection bias in sociological data. *American Sociological Review* 48:386-98.

Berk, Richard A., and S. C. Ray. 1982. Selection biases in sociological data. *Social Science Research* 11:351-98.

Bloom, Howard S. 1995. Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review* 19:547-56.

Bohrnstedt, George W. 1983. Measurement. In *Handbook of survey research*, edited by P. H. Rossi, James D. Wright, and Andy B. Anderson, 69-121. Orlando, FL: Academic Press.

Bonate, Peter L. 2000. *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall.

Boruch, Robert F. 1998. Randomized controlled experiments for evaluation. In *Handbook of applied social research methods*, edited by L. Bickman and Debra J. Rog, 161-90. Thousand Oaks, CA: Sage.

Burr, Jeffrey A., and John R. Nesselroade. 1990. Change measurement. In *Principles and structuring change*. Vol. 1, *Statistical methods in longitudinal research*, edited by A. von Eye, 3-35. Boston: Academic Press.

Cain, Kevin C., Richard A. Kronmal, and Andrzej S. Kosinski. 1992. Analysing the relationship between change in a risk factor and risk of disease. *Statistics in Medicine* 11:783-97.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Denver, CO: Houghton Mifflin.

Chambless, Lloyd E., and John R. Roeback. 1993. Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation. *Statistics in Medicine* 12:1213-37.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Dawson, Jeffrey D. 1998. Sample size calculations based on slopes and other summary statistics. *Biometrics* 54:323-30.

Donner, Allan. 1984. Approaches to sample size estimation in the design of clinical trials—A review. *Statistics in Medicine* 3:199-214.

Feldman, Henry A., and Sonja M. McKinlay. 1994. Cohort versus cross-sectional design in large field trials: Precision, sample size and a unifying model. *Statistics in Medicine* 13:61-78.

Feldt, Leonard S. 1957. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrica* 24:335-53.

Foster, E. Michael, and Leonard J. Bickman. 1996. An evaluator's guide to detecting attrition problems. *Evaluation Review* December:695-723.

Frison, Lars, and Stuart J. Pocock. 1992. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 11:1685-1704.

Fuller, W. A. 1987. *Measurement error models*. New York: John Wiley.

Greene, William H. 1997. *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Heckman, James J., and V. Joseph Hotz. 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84:862-77.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 91:945-60.

Hsieh, F. Y., Daniel A. Bloch, and Michael D. Larsen. 1998. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* 17:1623-34.

Kenny, D. A. 1975. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin* 82:345-62.

Kenny, D. A., and S. H. Cohen. 1979. A reexamination of selection and growth processes in the nonequivalent control group design. In *Sociological methodology 1980*, edited by K. Schuessler, 290-313. San Francisco: Jossey-Bass.

Lipsey, Mark. 1990. *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Lipsitz, Stuart R., and Michael Parzen. 1995. Sample size calculations for non-randomized trials. *The Statistician* 44:81-90.

Lord, Frederic M. 1960. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association* 55:307-21.

———. 1967. A paradox in the interpretation of group comparisons. *Psychological Bulletin* 68:304-5.

Maris, Eric. 1998. Covariance adjustment versus gain scores—Revisited. *Psychological Methods* 3:309-27.

McDowell, Ian, and Claire Newell. 1996. *Measuring health: A guide to rating scales and questionnaires*. New York: Oxford University Press.

Nunnally, Jum C., and Ira H. Bernstein. 1994. *Psychometric theory*. Vol. 3. New York: McGraw-Hill.

Pocock, Stuart J., and Diana R. Elbourne. 2000. Randomized trials or observational tribulations. *New England Journal of Medicine* 342:1907-9.

Pocock, Stuart J., Nancy L. Geller, and Anastasios A. Tsiatis. 1987. The analysis of multiple endpoints in clinical trials. *Biometrics* 43:487-98.

Reichardt, Charles S. 1979. *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Reichardt, Charles S., and Melvin M. Mark. 1998. Quasi-experimentation. In *Handbook of applied social research methods*, edited by L. Bickman and Debra J. Rog, 193-228. Thousand Oaks, CA: Sage.

Robins, Philip, and Daniel Friedlander. 1995. Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review* 85:39-52.

Rosenbaum, Paul R. 1995. *Observational studies*. New York: Springer-Verlag.

Rosner, Bernard. 1979. The analysis of longitudinal data in epidemiologic studies. *Journal of Chronic Disease* 32:163-73.

Rossi, Peter H., Howard E. Freeman, and Mark W. Lipsey. 1999. *Evaluation: A systematic approach*. 6th ed. Thousand Oaks, CA: Sage.

Rubin, Donald B. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2:1-26.

Rubin, Donald B. 1991. Practical implications of modes of statistical inference for causal effects and the critical role of random assignment. *Biometrics* 47:1213-34.

Self, Steven G., and Robert H. Mauritsen. 1988. Power/sample size calculations for generalized linear models. *Biometrics* 44:79-86.

Streiner, David L., and Geoffrey R. Norman. 1999. *Health measurement scales: A practical guide to their development and use*. New York: Oxford.

Werts, Charles E., and Robert L. Linn. 1970. A general linear model for studying growth. *Psychological Bulletin* 73:17-22.

Winship, Christopher, and Robert D. Mare. 1992. Models for sample selection bias. *Annual Review of Sociology* 18:327-50.

Winship, Christopher, and Stephen L. Morgan. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology* 25:659-707.

Yanez, N. David, Richard D. Kronmal, and Lynn R. Shemanski. 1998. The effects of measurement error in response variables and tests of association of explanatory variables in change models. *Statistics in Medicine* 17:2597-606.

Zar, J. H. 1984. *Biostatistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

J. Michael Oakes is a research scientist at New England Research Institutes, Inc. in Watertown, Massachusetts, where he focuses on research methods and social epidemiology. Current methodological interests include the measurement of SES, practical boundaries for endpoint distributions when censoring exceeds 50%, and using mixed-regression and geographic information system techniques to better understand neighborhood effects in community (i.e., cluster randomized) trials. Substantive interests include health disparities, cardiovascular disease prevention, and bioethics.

Henry A. Feldman is a principal research scientist at New England Research Institutes, Inc. and a lecturer in environmental health at Harvard University School of Public Health. His statistical work has included analytic methods in molecular and cellular biology, experimental studies of animal and human physiology, and community trials in cardiovascular health promotion.