

CHAPTER 2

Design Sensitivity

Statistical Power for Applied Experimental Research

Mark W. Lipsey

Sean M. Hurley

Appplied experimental research investigates the effects of deliberate intervention in situations of practical importance. A psychotherapist, for instance, might study the efficacy of systematic desensitization for reducing the symptoms of snake phobia, a school might evaluate the success of a drug education program, or a policymaker might ask for evidence that increasing the tax rate on gasoline will discourage consumption. The basic elements of experimental research are well-known: selection of participants and assignment of them to treatment and control conditions, preferably using a random procedure; application of the intervention of interest to the treatment group but not to the control group; monitoring the research situation to ensure that there are no differences between the treatment and control conditions other than the intervention; measurement of selected outcomes for both groups; and statistical analysis to determine if the groups differ on those dependent variable measures. To ensure that the conclusions about intervention effects drawn from experimental design are correct, the design must have both sensitivity and validity. *Sensitivity* refers to the likelihood that an effect, if present, will be detected. *Validity* refers to the likelihood that what is detected is, in fact, the effect of interest. This chapter is about the problem of sensitivity.

Sensitivity in intervention research is thus the ability to detect a difference between the treatment and control conditions on some outcome of interest. If the research design has high internal validity, that difference will represent the effect of

the intervention under investigation. What, then, determines our ability to detect it? Answering this question requires that we specify what is meant by *detecting a difference* in experimental research. Following current convention, we will take this to mean that statistical criteria are used to reject the null hypothesis of no difference between the mean on the outcome measure for the persons in the treatment condition and the mean for those in the control condition. In particular, we conclude that there is an effect if an appropriate statistical test indicates a statistically significant difference between the treatment and control means.

Our goal in this chapter is to help researchers “tune” experimental design to maximize sensitivity. However, before we can offer a close examination of the practical issues related to design sensitivity, we need to present a refined framework for describing and assessing the desired result—a high probability of detecting a given magnitude of effect if it exists. This brings us to the topic of *statistical power*, the concept that will provide the idiom for this discussion of design sensitivity.

The Statistical Power Framework

In the final analysis, applied experimental research comes down to just that: analysis (data analysis, that is). After all the planning, implementation, and data collection, the researcher is left with a set of numbers on which the crucial tests of statistical significance are conducted. There are four possible scenarios for this testing. There either *is* or *is not* a real treatment versus control difference that would be apparent if we had complete data for the entire population from which our sample was drawn (but we don't). And, for each of these situations, the statistical test on the sample data either *is* or *is not* significant. The various combinations can be depicted in a 2×2 table along with the associated probabilities, as shown in Table 2.1.

Finding statistical significance when, in fact, there is no effect is known as Type I error; the Greek letter α is used to represent the probability of that happening. Failure to find statistical significance when, in fact, there is an effect is known as Type II error; the Greek letter β is used to represent that probability. Most important, statistical power is the probability ($1 - \beta$) that statistical significance will be attained

Table 2.1 The Possibilities of Error in Statistical Significance Testing of Treatment (T) Versus Control (C) Group Differences

| Conclusion From Statistical Test on Sample Data | Population Circumstances | |
|--|---|--|
| | T and C Differ | T and C Do Not Differ |
| Significant difference (reject null hypothesis) | Correct conclusion Probability = $1 - \beta$ (power) | Type I error Probability = α |
| No significant difference (fail to reject null hypothesis) | Type II error Probability = β | Correct conclusion Probability = $1 - \alpha$ |

46 APPROACHES TO APPLIED RESEARCH

given that there really is an intervention effect. This is the probability that must be maximized for a research design to be sensitive to actual intervention effects.

Note that α and β in Table 2.1 are statements of *conditional* probabilities. They are of the following form: *If* the null hypothesis is true (false), *then* the probability of an erroneous statistical conclusion is α (β). When the null hypothesis is true, the probability of a statistical conclusion error is held to 5% by the convention of setting $\alpha = .05$. When the null hypothesis is false (i.e., there is a real effect), however, the probability of error is β , and β can be quite large. If we want to design experimental research in which statistical significance is found when the intervention has a real effect, then we must design for a low β error, that is, for high statistical power ($1 - \beta$).

An important question at this juncture concerns what criterion level of statistical power the researcher should strive for—that is, what level of risk for Type II error is acceptable? By convention, researchers generally set $\alpha = .05$ as the maximum acceptable probability of a Type I error. There is no analogous convention for beta. Cohen (1977, 1988) suggested $\beta = .20$ as a reasonable value for general use (more specifically, he suggested that power, equal to $1 - \beta$, be at least .80). This suggestion represents a judgment that Type I error is four times as serious as Type II error. This position may not be defensible for many areas of applied research where a null statistical result for a genuinely effective intervention may represent a great loss of valuable practical knowledge.

A more reasoned approach would be to analyze explicitly the cost-risk issues that apply to the particular research circumstances at hand (more on this later). At the first level of analysis, the researcher might compare the relative seriousness of Type I and Type II errors. If they are judged to be equally serious, the risk of each should be kept comparable; that is, alpha should equal beta. Alternatively, if one is judged to be more serious than the other, it should be held to a stricter standard even at the expense of relaxing the other. If a convention must be adopted, it may be wise to assume that, for intervention research of potential practical value, Type II error is at least as important as Type I error. In this case, we would set $\beta = .05$, as is usually done for α , and thus attempt to design research with power ($1 - \beta$) equal to .95.

Determinants of Statistical Power

There are four factors that determine statistical power: sample size, alpha level, statistical test, and effect size.

Sample Size. Statistical significance testing is concerned with sampling error, the expectable discrepancies between sample values and the corresponding population value for a given sample statistic such as a difference between means. Because sampling error is smaller for large samples, it is less likely to obscure real differences between means and statistical power is greater.

Alpha Level. The level set for alpha influences the likelihood of statistical significance—larger alpha makes significance easier to attain than does smaller alpha. When the null hypothesis is false, therefore, statistical power increases as alpha increases.

Statistical Test. Because investigation of statistical significance is made within the framework of a particular statistical test, the test itself is one of the factors determining statistical power.

Effect Size. If there is a real difference between the treatment and control conditions, the size of that difference will influence the likelihood of attaining statistical significance. The larger the effect, the more probable is statistical significance and the greater the statistical power. For a given dependent measure, effect size can be thought of simply as the difference between the means of the treatment versus control *populations*. In this form, however, its magnitude is partly a function of how the dependent measure is scaled. For most purposes, therefore, it is preferable to use an effect size formulation that standardizes differences between means by dividing by the standard deviation to adjust for arbitrary units of measurement. The effect size (*ES*) for a given difference between means, therefore, can be represented as follows:

$$ES = \frac{\mu_t - \mu_c}{\sigma}$$

where μ_t and μ_c are the respective means for the treatment and control populations and σ is their common standard deviation. This version of the effect size index was popularized by Cohen (1977, 1988) for purposes of statistical power analysis and is widely used in meta-analysis to represent the magnitude of intervention effects (Lipsey & Wilson, 2000). By convention, effect sizes are computed so that positive values indicate a “better” outcome for the treatment group than for the control group, and negative values indicate a “better” outcome for the control group.

For all but very esoteric applications, the most practical way actually to estimate the numerical values for statistical power is to use precomputed tables or a computer program. Particularly complete and usable reference works of statistical power tables have been published by Cohen (1977, 1988). Other general reference works along similar lines include those of Kraemer and Thiemann (1987), Lipsey (1990), and Murphy and Myors (2004). Among the computer programs available for conducting statistical power calculations are Power and Precision (from Biostat), nQuery Advisor (from Statistical Solutions), and SamplePower (from SPSS). In addition, there are open access power calculators on many statistical Web sites. The reader should turn to sources such as these for information on determining statistical power beyond the few illustrative cases presented in this chapter.

Figure 2.1 presents a statistical power chart for one of the more common situations. This chart assumes (a) that the statistical test used is a *t* test, one-way ANOVA, or other parametric test in this same family (more on this later) and (b) that the conventional $\alpha = .05$ level is used as the criterion for statistical significance. Given these circumstances, the chart shows the relationships among power ($1 - \beta$), effect size (*ES*), and sample size (*n* for each group) plotted on sideways log-log paper, which makes it easier to read values for the upper power levels and the lower

48 APPROACHES TO APPLIED RESEARCH

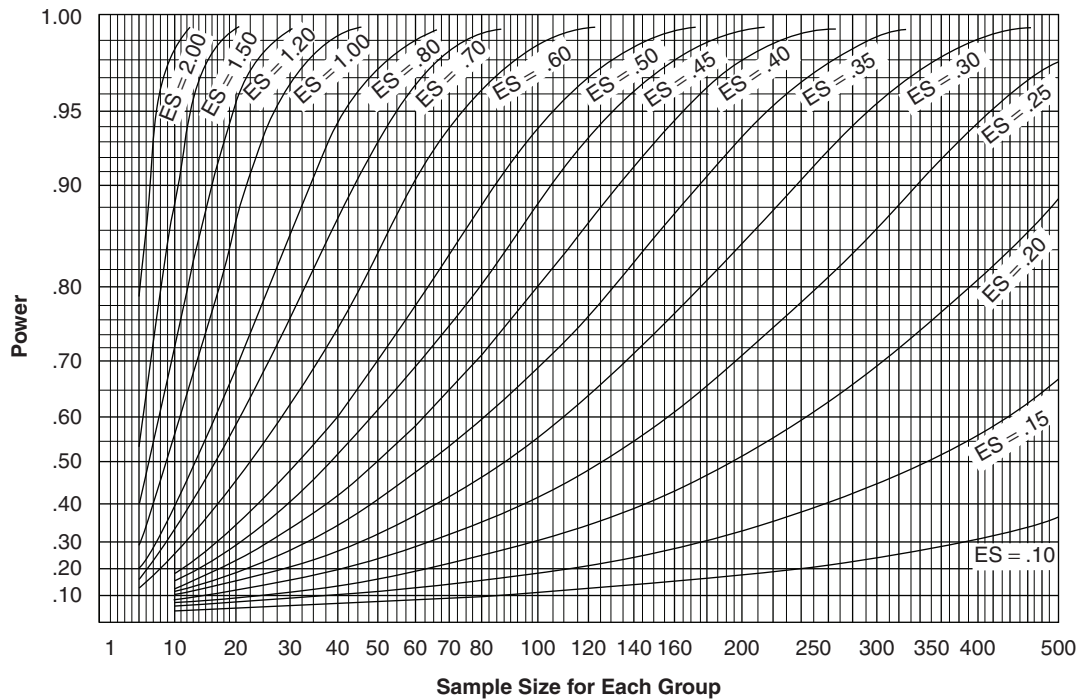


Figure 2.1 Power Chart for $\alpha = .05$, Two-Tailed, or $\alpha = .025$, One-Tailed

sample sizes. This chart shows, for instance, that if we have an experiment with 40 participants in each of the treatment and control groups (80 total), the power to detect an effect size of .80 (.8 standard deviations difference between the treatment and control group means) is about .94 (i.e., given a population $ES = .80$ and group $n = 40$, statistical significance would be expected 94% of the time at the $\alpha = .05$ level with a t test or one-way ANOVA).

Optimizing Statistical Power

To maximize the sensitivity of experimental research for detecting intervention effects using conventional criteria of statistical significance, the researcher must maximize statistical power. In the remainder of this chapter, we examine each of the determinants of statistical power and discuss how it can be manipulated to enhance power. The objective of this discussion is to provide the researcher with the conceptual tools to design experimental research with the greatest possible sensitivity to intervention effects given the resources available. Moreover, in those cases where an appropriately high level of statistical power cannot be attained, these same concepts can be used to analyze the limitations of the research design and guard against misinterpretation.

Sample Size

The relationship between sample size and statistical power is so close that many textbooks discuss power only in terms of determining the sample size necessary to attain a desired power level. A look at Figure 2.1 makes clear why sample size warrants so much attention. Virtually any desired level of power for detecting any given effect size can be attained by making the samples large enough.

The difficulty that the relationship between sample size and statistical power poses for intervention research is that the availability of participants is often limited. Although a researcher can increase power considerably by parading a larger number of participants through the study, there must be individuals ready to march before this becomes a practical strategy. In practical intervention situations, relatively few persons may be appropriate for the intervention or, if there are enough appropriate persons, there may be limits on the facilities for treating them. If facilities are adequate, there may be few who volunteer or whom program personnel are willing to assign; or, if assigned, few may sustain their participation until the study is complete. The challenge for the intervention researcher, therefore, is often one of keeping power at an adequate level with modest sample sizes. If modest sample sizes in fact generally provided adequate power, this particular challenge would not be very demanding. Unfortunately, they do not.

Suppose, for instance, that we decide that $ES = .20$ is the minimal effect size that we would want our intervention study to be able to detect reliably. An ES of $.20$ is equivalent to a 22% improvement in the success rate for the treatment group (more on this later). It is also the level representing the first quintile in the effect size distribution derived from meta-analyses of psychological, behavioral, and education intervention research (Lipsey & Wilson, 1993). Absent other considerations, therefore, $ES = .20$ is a reasonable minimal effect size to ask research to detect—it is not so large that it requires heroic assumptions to think it might actually be produced by an intervention and not so small that it would clearly lack practical significance.

If we calculate the sample size needed to yield a power level of $.95$ ($\beta = \alpha = .05$), we find that the treatment and control group must each have a minimum of about 650 participants for a total of about 1,300 in both groups (see Figure 2.1). The sample sizes in social intervention research are typically much smaller than that, often less than 100 in each group. If we want to attain a power level for $ES = .20$ that makes Type II error as small as the conventional limit on Type I error through sample size alone, then we must increase the number of participants quite substantially over the average in present practice. Even attaining the more modest $.80$ power level suggested as a minimum by Cohen (1988) would require a sample size of about 400 per treatment group, larger than many studies can obtain.

Increased sample size is thus an effective way to boost statistical power and should be employed whenever feasible, but its costs and limited availability of participants may restrict the researcher's ability to use this approach. It is important, therefore, that the researcher be aware of other routes to increasing statistical power. The remainder of this chapter discusses some of these alternate routes.

Alpha Level

Alpha is conventionally set at .05 for statistical significance testing and, on the surface, may seem to be the one straightforward and unproblematic element of statistical power for the intervention researcher. That impression is misleading. An α of .05 corresponds to a .95 probability of a correct statistical conclusion only when the null hypothesis is true. However, a relatively conservative alpha makes statistical significance harder to attain when the null hypothesis is false and, therefore, decreases the statistical power. Conversely, relaxing the alpha level required for statistical significance increases power. The problem is that this reduction in the probability of a Type II error comes at the expense of an increased probability of a Type I error. This means that the researcher cannot simply raise alpha until adequate power is attained but, rather, must find some appropriate balance between alpha and beta. Both Type I error (α) and Type II error (β) generally have important implications in the investigation of intervention effects. Type I error can mean that an ineffective or innocuous intervention is judged beneficial or, possibly, harmful, whereas Type II error can permit a truly effective intervention (or a truly harmful one) to go undiscovered. Though little has been written in recent years about how to think about this balancing act, useful perspectives can be found in Brown (1983), Cascio and Zedeck (1983), Nagel and Neef (1977), and Schneider and Darcy (1984). In summary form, the advice of these authors is to consider the following points in setting error risk levels.

Prior Probability. Because the null hypothesis is either true or false, only one type of inferential error is possible in a given study—Type I for a true null hypothesis and Type II for a false null hypothesis. The problem, of course, is that we do not know if the null hypothesis is true or false and, thus, do not know which type of error is relevant to our situation. However, when there is evidence that makes one alternative more likely, the associated error should be given more importance. If, for example, prior research tends to show an intervention effect, the researcher should be especially concerned about protection against Type II error and should set beta accordingly.

Directionality of Significance Testing. A significance test of a one-tailed hypothesis (e.g., that the treatment group mean is superior to the control group) conducted at a given α level has higher power (smaller beta) than a two-tailed test at the same alpha (e.g., that the treatment group is either superior *or* inferior to control). In applied intervention research, concern often centers on one direction of effects, for instance, whether a new intervention is better than an existing one. In these situations, it may be reasonable to argue that one-tailed tests are justified and that using two-tailed tests amounts to inappropriate restriction of the alpha level. Such an argument implies that a negative intervention effect, should it occur, is of no interest, however—a rather strong claim for many kinds of intervention.

Relative Costs and Benefits. Perhaps the most important aspect of error risk in intervention research has to do with the consequences of an error. Rarely will the costs of each type of error be the same, nor will the benefits of each type of correct inference. Sometimes, intervention effects and their absence can be interpreted directly in

terms of dollars saved or spent, lives saved or lost, and the like. In such cases, the optimal relationship between alpha and beta error risk should be worked out according to their relative costs and benefits. When the consequences of Type I and Type II errors cannot be specified in such definite terms, the researcher may still be able to rely on some judgment about the relative seriousness of the risks. Such judgment might be obtained by asking those familiar with the intervention circumstances to rate the error risk and the degree of certainty that they feel is minimal for the conclusions of the research. This questioning, for instance, may reveal that knowledgeable persons believe, on average, that a 95% probability of detecting a meaningful effect is minimal and that Type II error is three times as serious as Type I error. This indicates that β should be set at .05 and α at .15. Nagel and Neef (1977) provided a useful decision theory approach to this judgment process that has the advantage of requiring relatively simple judgments from those whose views are relevant to the research context.

If some rational analysis of the consequences of error is not feasible, it may be necessary to resort to a convention (such as $\alpha = .05$) as a default alternative. For practical intervention research, the situation is generally one in which *both* types of errors are serious. Under these circumstances, the most straightforward approach is to set alpha risk and beta risk equal unless there is a clear reason to do otherwise. If we hold to the usual convention that α should be .05, then we should design research so that β will also be .05. If such high standards are not practical, then both alpha and beta could be relaxed to some less stringent level—for example, .10 or even .20.

To provide some framework for consideration of the design issues related to the criterion levels of alpha and beta set by the researcher, Table 2.2 shows the required sample size per group for the basic two-group experimental design at various effect sizes under various equal levels of alpha (two-tailed) and beta. It is noteworthy that maintaining relatively low levels of alpha and beta risk (e.g., .05 or below) requires either rather large effect sizes or rather large sample sizes. Moreover, relaxing alpha levels does not generally yield dramatic increases in statistical power for the most difficult to detect effect sizes. Manipulation of other aspects of the power function, such as those described later, will usually be more productive for the researcher seeking to detect potentially modest effects with modest samples sizes.

Statistical Test

Consider the prototypical experimental design in which one treatment group is compared with one control group. The basic statistical tests for analyzing this design are the familiar *t* test and one-way analysis of variance (ANOVA). These tests use an “error term” based on the within-group variability in the sample data to assess the likelihood that the mean difference between the groups could result from sampling error. To the extent that within-group variability can be eliminated, minimized, or somehow offset, intervention research will be more powerful—that is, more sensitive to true effects if they are present.

Two aspects of the statistical test are paramount in this regard. First, for a given set of treatment versus control group data, different tests may have different formulations

52 APPROACHES TO APPLIED RESEARCH

Table 2.2 Approximate Sample Size for Each Group Needed to Attain Various Equal Levels of Alpha and Beta for a Range of Effect Sizes

| <i>Effect Size</i> | <i>Level of Alpha and Beta ($\alpha = \beta$)</i> | | | |
|--------------------|--|------------|------------|------------|
| | <i>.20</i> | <i>.10</i> | <i>.05</i> | <i>.01</i> |
| .10 | 900 | 1,715 | 2,600 | 4,810 |
| .20 | 225 | 430 | 650 | 1,200 |
| .30 | 100 | 190 | 290 | 535 |
| .40 | 60 | 110 | 165 | 300 |
| .50 | 35 | 70 | 105 | 195 |
| .60 | 25 | 50 | 75 | 135 |
| .70 | 20 | 35 | 55 | 100 |
| .80 | 15 | 30 | 45 | 75 |
| .90 | 10 | 25 | 35 | 60 |
| 1.00 | 10 | 20 | 30 | 50 |

of the sampling error estimate and the critical test values needed for significance. For instance, nonparametric tests—those that use only rank order or categorical information from dependent variable scores—generally have less inherent power than do parametric tests, which use scores representing degrees of the variable along some continuum.

The second and most important aspect of a statistical test that is relevant to power is the way it partitions sampling error and which components of that error variance are used in the significance test. It is often the case in intervention research that some of the variability on a given dependent measure is associated with participant characteristics that are not likely to change as a result of intervention. If certain factors extraneous to the intervention effect of interest contribute to the population variability on the dependent measure, the variability associated with those factors can be removed from the estimate of sampling error against which differences between treatment and control means are tested with corresponding increases in power.

A simple example might best illustrate the issue. Suppose that men and women, on average, differ in the amount of weight they can lift. Suppose further that we want to assess the effects of an exercise regimen that is expected to increase muscular strength. Forming treatment and control groups by simple random sampling of the undifferentiated population would mean that part of the within-group variability that is presumed to reflect the luck of the draw (sampling error) would be the natural differences between men and women. This source of variability may well be judged irrelevant to an assessment of the intervention effect—the intervention may rightfully be judged effective if it increases the strength of women relative to the natural variability in women's strength and that of men relative to the natural variability in men's strength. The corresponding sampling procedure is not

simple random sampling but stratified random sampling, drawing women and men separately so that the experimental sample contains identified subgroups of women and men. The estimate of sampling error in this case comes from the within-group variance—within experimental condition within gender—and omits the between-gender variance, which has now been identified as having a source other than the luck of the draw.

All statistical significance tests assess effects relative to an estimate of sampling error but they may make different assumptions about the nature of the sampling and, hence, the magnitude of the sampling error. The challenge to the intervention researcher is to identify the measurable extraneous factors that contribute to population variability and then use (or assume) a sampling strategy and corresponding statistical test that assesses intervention effects against an appropriate estimate of sampling error. Where there are important extraneous factors that correlate with the dependent variable (and there almost always are), using a statistical significance test that partitions them out of the error term can greatly increase statistical power. With this in mind, we review below some of the more useful of the variance control statistical designs with regard to their influence on power.

Analysis of Covariance

One of the most useful of the variance control designs for intervention research is the one-way analysis of covariance (ANCOVA). Functionally, the ANCOVA is like the simple one-way ANOVA, except that the dependent variable variance that is correlated with a covariate variable (or linear combination of covariate variables) is removed from the error term used for significance testing. For example, a researcher with a reading achievement test as a dependent variable may wish to remove the component of performance associated with IQ before comparing the treatment and control groups. IQ differences may well be viewed as nuisance variance that is correlated with reading scores but is not especially relevant to the impact of the program on those scores. That is, irrespective of a student's IQ score, we would still expect an effective reading program to boost the reading score.

It is convenient to think of the influence of variance control statistical designs on statistical power as a matter of adjusting the effect size in the power relationship. Recall that ES , as it is used in statistical power determination, is defined as $(\mu_t - \mu_c)/\sigma$ where σ is the pooled within-groups standard deviation. For assessing the power of variance control designs, we adjust this ES to create a new value that is the one that is operative for statistical power determination. For the ANCOVA statistical design, the operative ES for power determination is as follows:

$$ES_{ac} = \frac{\mu_t - \mu_c}{\sigma\sqrt{1 - r_{dc}^2}},$$

where ES_{ac} is the effect size formulation for the one-way ANCOVA; μ_t and μ_c are the means for the treatment and control populations, respectively; σ is the common

54 APPROACHES TO APPLIED RESEARCH

standard deviation; and r_{dc} is the correlation between the dependent variable and the covariate. As this formula shows, the operative effect size for power determination using ANCOVA is inflated by a factor of $1/\sqrt{1-r^2}$, which multiplies ES by 1.15 when $r = .50$, and 2.29 when $r = .90$. Thus, when the correlation of the covariate(s) with the dependent variable is substantial, the effect of ANCOVA on statistical power can be equivalent to more than doubling the operative effect size. Examination on Figure 2.1 reveals that such an increase in the operative effect size can greatly enhance power at any given sample size.

An especially useful application of ANCOVA in intervention research is when both pretest and posttest values on the dependent measure are available. In many cases of experimental research, preexisting individual differences on the characteristic that intervention is intended to change will not constitute an appropriate standard for judging intervention effects. Of more relevance will be the size of the intervention effect relative to the dispersion of scores for respondents that began at the same initial or baseline level on that characteristic. In such situations, a pretest measure is an obvious candidate for use as a covariate in ANCOVA. Because pretest-posttest correlations are generally high, often approaching the test-retest reliability of the measure, the pretest as a covariate can dramatically increase the operative effect size in statistical power. Indeed, ANCOVA with the pretest as the covariate is so powerful and so readily attainable in most instances of intervention research that it should be taken as the standard to be used routinely unless there are good reasons to the contrary.

ANOVA With a Blocking Factor

In the blocked ANOVA design, participants are first categorized into blocks, that is, groups of participants who are similar to each other on some characteristic related to the dependent variable. For example, to use gender as a blocking variable, one would first divide participants into males and females, then assign some males to the treatment group and the rest to the control group and, separately, assign some females to treatment and the rest to control.

In the blocked design, the overall variance on the dependent measure can be viewed as the sum of two components: the within-blocks variance and the between-blocks variance. Enhanced statistical power is gained in this design because it removes the contribution of the between-blocks variance from the error term against which effects are tested. As in the ANCOVA case, this influence on power can be represented in terms of an adjusted effect size. If we let PV_b equal the proportion of the total dependent variable variance associated with the difference between blocks, the operative ES for this case is as follows:

$$ES_{ab} = \frac{\mu_t - \mu_c}{\sigma\sqrt{1 - PV_b}},$$

where ES_{ab} is the effect size formulation for the blocked one-way ANOVA, σ is the pooled within-groups standard deviation (as in the unadjusted ES), and PV_b is

σ_b^2/σ^2 with σ_b^2 the between-blocks variance and σ^2 the common variance of the treatment and control populations.

The researcher, therefore, can estimate PV_b , the between-blocks variance, as a proportion of the common (or pooled) variance within experimental groups and use it to adjust the effect size estimate in such a way as to yield the operative effect size associated with the statistical power of this design. If, for instance, the blocking factor accounts for as much as half of the common variance, the operative *ES* increases by more than 40%, with a correspondingly large increase in power.

Power Advantages of Variance Control Designs

The variance control statistical designs described above all have the effect of reducing the denominator of the effect size index and, hence, increasing the operative effect size that determines statistical power. Depending on the amount of variance controlled in these designs, the multiplier effect on the effect size can be quite considerable. Table 2.3 summarizes that multiplier effect for different proportions of the within-groups variance associated with the control variable. Although the effects are modest when the control variable accounts for a small proportion of the dependent variable variance, they are quite considerable for higher proportions. For instance, when the control variable accounts for as much as 75% of the variance, the operative effect size is double what it would be without the control variable. Reference back to Figure 2.1, the statistical power chart, will reveal that a doubling of the effect size has a major effect on statistical power. Careful use of variance control designs, therefore, is one of the most important tactics that the intervention researcher can use to increase statistical power without requiring additional participants in the samples.

Effect Size

The effect size parameter in statistical power can be thought of as a signal-to-noise ratio. The signal is the difference between treatment and control population means on the dependent measure (the *ES* numerator, $\mu_t - \mu_c$). The noise is the within-groups variability on that dependent measure (the *ES* denominator, σ). Effect size and, hence, statistical power is large when the signal-to-noise ratio is high—that is, when the *ES* numerator is large relative to the *ES* denominator. In the preceding section, we saw that variance control statistical designs increase statistical power by removing some portion of nuisance variance from the *ES* denominator and making the operative *ES* for statistical power purposes proportionately larger. Here, we will look at some other approaches to increasing the signal-to-noise ratio represented by the effect size.

Dependent Measures

The dependent measures in intervention research yield the set of numerical values on which statistical significance testing is performed. Each such measure chosen

56 APPROACHES TO APPLIED RESEARCH

Table 2.3 Multiplier by Which *ES* Increases When a Covariate or Blocking Variable Is Used to Reduce Within-Groups Variance

| <i>Proportion of Variance Associated With Control Variable^a</i> | <i>Multiplier for ES Increase</i> |
|--|-----------------------------------|
| .05 | 1.03 |
| .10 | 1.05 |
| .15 | 1.08 |
| .20 | 1.12 |
| .25 | 1.15 |
| .30 | 1.20 |
| .35 | 1.24 |
| .40 | 1.29 |
| .45 | 1.35 |
| .50 | 1.41 |
| .55 | 1.49 |
| .60 | 1.58 |
| .65 | 1.69 |
| .70 | 1.83 |
| .75 | 2.00 |
| .80 | 2.24 |
| .85 | 2.58 |
| .90 | 3.16 |
| .95 | 4.47 |
| .99 | 10.00 |

a. r^2 for ANCOVA, PV_b for blocked ANOVA.

for a study constitutes a sort of listening station for certain effects expected to result from the intervention. If the listening station is in the wrong place or is unresponsive to effects when they are actually present, nothing will be heard. To optimize the signal-to-noise ratio represented in the effect size, the ideal measure for intervention effects is one that is maximally responsive to any change that the intervention brings about (making a large *ES* numerator) and minimally responsive to anything else (making a small *ES* denominator). In particular, three aspects of outcome measurement have direct consequences for the magnitude of the effect size parameter and, therefore, statistical power: (a) validity for measuring change, (b) reliability, and (c) discrimination of individual differences among respondents.

Validity for Change. For a measure to respond to the signal, that is, to intervention effects, it must, of course, be a valid measure of the characteristic that the intervention is expected to change. But validity alone is not sufficient to make a measure responsive to intervention effects. What is required is validity for *change*. A measure can be a valid indicator of a characteristic but still not be a valid indicator of change on that characteristic. Validity for change means that the measure shows an observable *difference* when there is, in fact, a change on the characteristic measured that is of sufficient magnitude to be interesting in the context of application.

There are various ways in which a measure can lack validity for change. For one, it may be scaled in units that are too gross to detect the change. A measure of mortality (death rate), for instance, is a valid indicator of health status but is insensitive to variations in how sick people are. Graduated measures, those that range over some continuum, are generally more sensitive to change than categorical measures, because the latter record changes only between categories, not within them. The number of readmissions to a mental hospital, for example, constitutes a continuum that can differentiate one readmission from many. This continuum is often represented categorically as “readmitted” versus “not readmitted,” however, with a consequent loss of sensitivity to change and statistical power.

Another way in which a measure may lack validity for measuring change is by having a floor or ceiling that limits downward or upward response. A high school-level mathematics achievement test might be quite unresponsive to improvements in Albert Einstein’s understanding of mathematics—he would most likely score at the top of the scale with or without such improvements. Also, a measure may be specifically designed to cancel out certain types of change, as when scores on IQ tests are scaled by age norms to adjust away age differences in ability to answer the items correctly.

In short, measures that are valid for change will respond when intervention alters the characteristic of interest and, therefore, will differentiate a treatment group from a control group. The stronger this differentiation, the greater the contrast between the group means will be and, correspondingly, the larger the effect size.

Reliability. Turning now to the noise in the signal detection analogy, we must consider variance in the dependent measure scores that may obscure any signal due to intervention effects. Random error variance—that is, unreliability in the measure—is obviously such a noise. Unreliability represents fluctuations in the measure that are unrelated to the characteristic being measured, including intervention effects on that characteristic. Measures with lower measurement error will yield less variation in the distribution of scores for participants within experimental groups. Because within-groups variance is the basis for the denominator of the *ES* ratio, less measurement error makes that denominator smaller and the overall *ES* larger.

Some measurement error is intrinsic—it follows from the properties of the measure. Self-administered questionnaires, for instance, are influenced by fluctuations in respondents’ attention, motivation, comprehension, and so forth. Some measurement error is procedural—it results from inconsistent or inappropriate application of the measure. Raters who must report on an observed characteristic,

58 APPROACHES TO APPLIED RESEARCH

for instance, may not be trained to use the same standards for their judgment, or the conditions of observation may vary for different study participants in ways that influence their ratings.

Also included in measurement error is systematic but irrelevant variation—response of the measure to characteristics other than the one of interest. When these other characteristics vary differently than the one being measured, they introduce noise into a measure. For example, frequency of arrest, which may be used to assess the effects of intervention for juvenile delinquency, indexes police behavior (e.g., patrol and arrest practices) as well as the criminal behavior of the juveniles. If the irrelevant characteristic to which the measure is also responding can be identified and separately measured, its influence can be removed by including it as a covariate in an ANCOVA, as discussed above. For instance, if we knew the police precinct in which each arrest was made, we could include that information as control variables (dummy coding each precinct as involved vs. not involved in a given arrest) that would eliminate variation in police behavior across precincts from the effect size for a delinquency intervention.

Discrimination of Individual Differences. Another source of systematic but often irrelevant variation that is especially important in intervention effectiveness research has to do with relatively stable individual differences on the characteristic measured. When a measure is able to discriminate strongly among respondents, the variance of its distribution of scores is increased. This variation does not represent error, as respondents may truly differ, but it nonetheless contributes to the noise variance that can obscure intervention effects. In a reading improvement program, for example, the primary interest is whether each participant shows improvement in reading level, irrespective of his or her initial reading level, reading aptitude, and so forth. If the measure selected is responsive to such other differences, the variability may be so great as to overshadow any gains from the program.

Where psychological and educational effects of intervention are at issue, an important distinction is between “psychometric” measures, designed primarily to discriminate individual differences, and “edumetric” measures, designed primarily to detect change (Carver, 1974). Psychometric measures are those developed using techniques that spread out the scores of respondents; IQ tests, aptitude tests, personality tests, and other such standardized tests would generally be psychometric measures. By comparison, edumetric measures are those developed through the sampling of some defined content domain that represents the new responses participants are expected to acquire as a result of intervention. Mastery tests, such as those an elementary school teacher would give students to determine whether they have learned to do long division, are examples of edumetric tests.

Because they are keyed specifically to the sets of responses expected to result from intervention, edumetric tests, or measures constructed along similar lines, are more sensitive than psychometric tests to the changes induced by intervention and less sensitive to preexisting individual differences. To the extent that any measure reflects less heterogeneity among participants, within-group variability on that measure is smaller. That, in turn, results in a smaller denominator for the *ES* ratio and a corresponding increase in statistical power.

The Independent Variable

The independent variable in intervention research is defined by the contrast between the experimental conditions (e.g., treatment and control) to which participants are exposed. When more contrast is designed into the study, the effect size can be correspondingly larger if the intervention is effective.

Dose Response. Experimental design is based on the premise that intervention levels can be made to vary and that different levels might result in different responses. Generally speaking, the “stronger” the intervention, the larger the response should be. One way to attain a large effect size, therefore, is to design intervention research with the strongest possible dose of the intervention represented in the treatment condition. In testing a new math curriculum, for instance, the researcher might want the teachers to be very well-trained to deliver it and to spend a significant amount of class time doing so. If the intervention is effective, the larger effect size resulting from a stronger dose will increase statistical power for detecting the effect.

Optimizing the strength of the intervention operationalized in research requires some basis for judging what might constitute the optimal configuration for producing the expected effects. There may be insufficient research directly on the intervention under study (else why do the research), but there may be other sources of information that can be used to configure the intervention so that it is sufficiently strong to potentially show detectable effects. One source, for example, is the experience and intuition of practitioners in the domain where the intervention, or variants, is applied.

Variable Delivery of the Intervention. The integrity or fidelity of an intervention is the degree to which it is delivered as planned and, in particular, the degree to which it is delivered in a uniform manner in the right amounts to the right participants at the right time. At one end of the continuum, we might consider the case of intervention research conducted under tightly controlled clinical or laboratory conditions in which delivery can be regulated very closely. Under these conditions, we would expect a high degree of intervention integrity, that is, delivery of a constant, appropriate dose to each participant.

Intervention research, however, cannot always be conducted under such carefully regulated circumstances. It must often be done in the field with volunteer participants whose compliance with the intervention regimen is difficult to ensure. Moreover, the interventions of interest are often not those for which dosage is easily determined and monitored, nor are they necessarily delivered uniformly. The result is that the participants in a treatment group may receive widely different amounts and even kinds of intervention (e.g., different mixes of components). If participants’ responses to intervention vary with its amount and kind, then it follows that variation in the intervention will generate additional variation in the outcome measures.

When treatment and control groups are compared in a statistical analysis, all that usually registers as an intervention effect is the difference between the treatment group’s mean score and the control group’s mean score on the dependent

variable. If there is variation around those means, it goes into the within-groups variance of the effect size denominator, making the overall *ES* smaller. Maintaining a uniform application of treatment and control conditions is the best way to prevent this problem. One useful safeguard is for the researcher to actually measure the amount of intervention received by each participant in the treatment and control conditions (presumably little or none in the control). This technique yields information about how much variability there actually was and generates a covariate that may permit statistical adjustment of any unwanted variability.

Control Group Contrast. Not all aspects of the relationship between the independent variable and the effect size have to do primarily with the intervention. The choice of a control condition also plays an important role. The contrast between the treatment and control means can be heightened or diminished by the choice of a control that is more or less different from the treatment condition in its expected effects on the dependent measure.

Generally, the sharpest contrast can be expected when what the control group receives involves no aspects of the intervention or any other attention—that is, a “no treatment” control. For some situations, however, this type of control may be unrepresentative of participants’ experiences in nonexperimental conditions or may be unethical. This occurs particularly for interventions that address problems that do not normally go unattended—severe illness, for example. In such situations, other forms of control groups are often used. The “treatment as usual” control group, for instance, receives the usual services in comparison to a treatment group that receives innovative services. Or a placebo control might be used in which the control group receives attention similar to that received by the treatment group but without the specific active ingredient that is presumed to be the basis of the intervention’s efficacy. Finally, the intervention of interest may simply be compared with some alternative intervention, for example, traditional psychotherapy compared with behavior modification as treatment for anxiety.

The types of control conditions described above are listed in approximate order according to the magnitude of the contrast they would generally be expected to show when compared with an effective intervention. The researcher’s choice of a control group, therefore, will influence the size of the potential contrast and hence of the potential effect size that appears in a study. Selection of the control group likely to show the greatest contrast from among those appropriate to the research issues can thus have an important bearing on the statistical power of the design.

Statistical Power for Multilevel Designs

For the experimental designs discussed in the previous sections, we have assumed that the units on which the dependent variables were measured are the same units that were randomly assigned to treatment and control conditions. In social science intervention studies, those units are typically individual people. Research designs

for some intervention situations, however, involve assignment of clusters of units to experimental conditions or delivery of treatment at the cluster level, but measurement of the outcomes on the individual units within those clusters. Such designs are especially common in education research where classrooms or entire schools may be assigned to treatment and control conditions with student grades or achievement test scores as the dependent variable. Similarly, patients whose outcomes are of interest might be clustered within hospitals assigned to treatment and control conditions, energy use might be examined for apartments clustered within housing projects assigned to receive a weatherization program or not, and so forth. Even when individuals are randomly assigned to conditions, if the treatment and control conditions are implemented on clusters, for example, classrooms, there are still multiple levels in the design. These types of designs may also have other levels or groupings in between the units of measurement and the units of randomization. For example, students (whose achievement scores are the outcomes of interest) might be clustered within classrooms that are clustered within schools that are clustered within school districts that are assigned to intervention and control conditions. For simplicity, the discussion here will be limited to two-level models, but the general principles can be extended to designs with more than two levels.

These cluster or multilevel designs have distinct characteristics that affect statistical power. One way to think about them is in terms of the sample size for the experiment—a critical factor for power discussed earlier. Is the pertinent sample size the number of clusters assigned to the experimental conditions or is it the number of units within all those clusters on which the outcomes are measured? The answer, and the main source of complexity for power analysis, is that it could be either or something in between. The operative sample size is the number of *statistically independent* units represented in the study. Participants within a cluster (e.g., students within a classroom) are likely to have dependent measure scores that are more similar to each other than to participants in different clusters either because of the natural sorting processes that have put them in that cluster or because of similar influences that they share as members of it. If so, their scores are not statistically independent—there is some degree of predictability from one to another within a classroom. When there is statistical dependence among the scores within clusters, the operative sample size is no longer the number of units measured but, instead, shrinks toward the number of clusters assigned, which is always a smaller number (Snijders & Bosker, 1999).

Statistical analysis for multilevel designs and, correspondingly, statistical power considerations must, therefore, take into account the within- and between-cluster variance structure of the data. If there is relative homogeneity within clusters and heterogeneity between clusters, the results will be quite different than if it is the other way around. Specialized statistical programs are available for analyzing multilevel data, for example, HLM (Raudenbush, Bryk, & Congdon, 2004), MLwiN (Rasbash, Steele, Browne, & Prosser, 2004), and, more generally, mixed models analysis routines in the major computer programs such as SPSS, SAS, and Stata. In the sections that follow, we identify the distinctive issues associated with statistical power in multilevel designs and describe ways in which it can be optimized and estimated.

Determinants of Statistical Power for Multilevel Designs

Basically, the same four factors that influence power in single-level designs apply to multilevel designs—sample size, alpha level, the statistical test (especially whether variance controls are included), and effect size. The alpha level at which the intervention effect is tested and the effect size are defined virtually the same way in multilevel designs as in single-level ones and function the same way in power analysis. It should be particularly noted that despite the greater complexity of the structure of the variance within treatment and control groups in multilevel designs, the effect size parameter remains the same. It is still defined as the difference between the mean score on the dependent variable for all the individuals in the treatment group and the mean for all the individuals in the control group divided by the common standard deviation of all the scores within the treatment and control groups. In a multilevel design, the variance represented in that standard deviation could, in turn, be decomposed into between- and within-cluster components or built up from them. It is, nonetheless, the same treatment or control population variance (estimated from sample values) irrespective of whether the participants providing scores have been sampled individually or clusterwise.

The statistical analysis on the other hand will be different—it will involve a multilevel statistical model that represents participant scores at the lowest level and the clusters that were randomized at the highest level. One important implication of this multilevel structure is that variance control techniques, such as use of selected covariates, can be applied at both the participant and cluster levels of the analysis. Similarly, sample size applies at both levels and involves the number of clusters assigned to experimental conditions and the number of participants within clusters who provide scores on the dependent measures.

One additional factor distinctive to multilevel designs also plays an important role in statistical power: the intraclass correlation (ICC; Hox, 2002; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The ICC is a measure of the proportion of the total variance of the dependent variable scores that occurs between clusters. It can be represented as follows:

$$\rho = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{between}}^2 + \sigma_{\text{within}}^2},$$

where the numerator is the variance between the clusters and the denominator is the total variance in the model (between-cluster plus within-cluster variance).

If none of the variability in the data is accounted for by between-cluster differences, then the ICC will be 0 and the effective sample size for the study will simply be the total number of participants in the study. If, on the other hand, all the variability is accounted for by between-cluster differences, then the ICC will be 1 and the effective N for the study will be the number of clusters. In practice, the ICC will be somewhere between these two extremes, and the effective N of the study will be somewhere in between the number of participants and the number of clusters.

Figure 2.2 contains a graph that depicts the effect of the magnitude of the ICC on the power to detect an effect size of .40 at $\alpha = .05$ with 50 clusters total (evenly divided between treatment and control) and 15 participants per cluster. As the figure shows, even small increases in the ICC can substantially reduce the power.

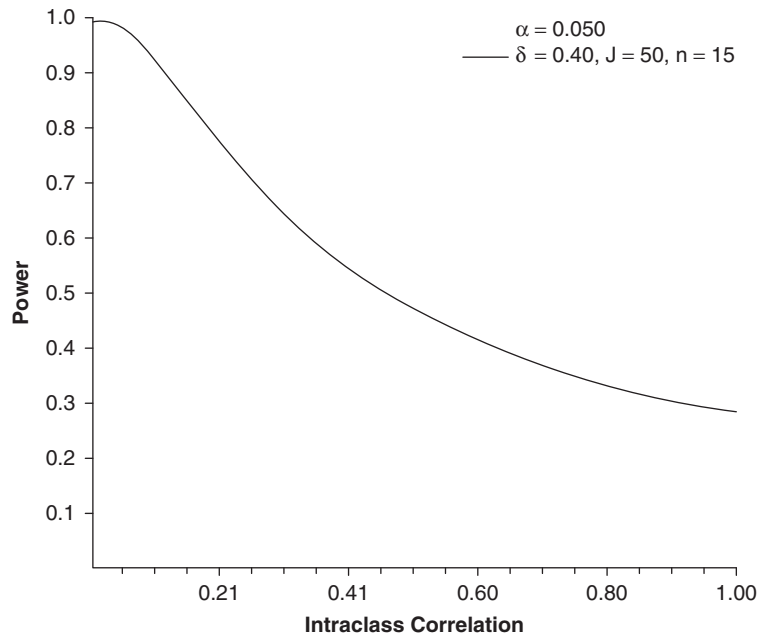


Figure 2.2 The Relationship Between ICC and Power to Detect an Effect Size of .40, With 50 Clusters Total, 15 Participants per Cluster, and $\alpha = .05$ (graph generated using optimal design software)

Clearly, the ICC is crucial for determining statistical power when planning a study. Unfortunately, the researcher has no control over what the ICC will be for a particular study. Thus, when estimating the statistical power of a planned study, the researcher should consider the ICC values that have been reported for similar research designs. For example, the ICCs for the educational achievement outcomes of students clustered within classroom or schools typically range from approximately .15 to .25 (Hedges & Hedberg, 2006).

Unlike the ICC, the number of clusters and the number of participants within each cluster are usually within the researcher's control, at least to the extent that resources allow. Unfortunately, in multilevel analyses the total number of participants (which are usually more plentiful) has less of an effect on power than the number of clusters (which are often available only in limited numbers). This is in contrast to single-level designs in which the sample size at the participant level plays a large role in determining power. See Figure 2.3 for a graph depicting the relationship between sample size at the participant level and power to detect an effect size of .40 at $\alpha = .05$ for a study with 50 clusters total and an ICC of .20. Once clusters have around 15 participants each, adding additional participants yields only modest gains in power.

64 APPROACHES TO APPLIED RESEARCH

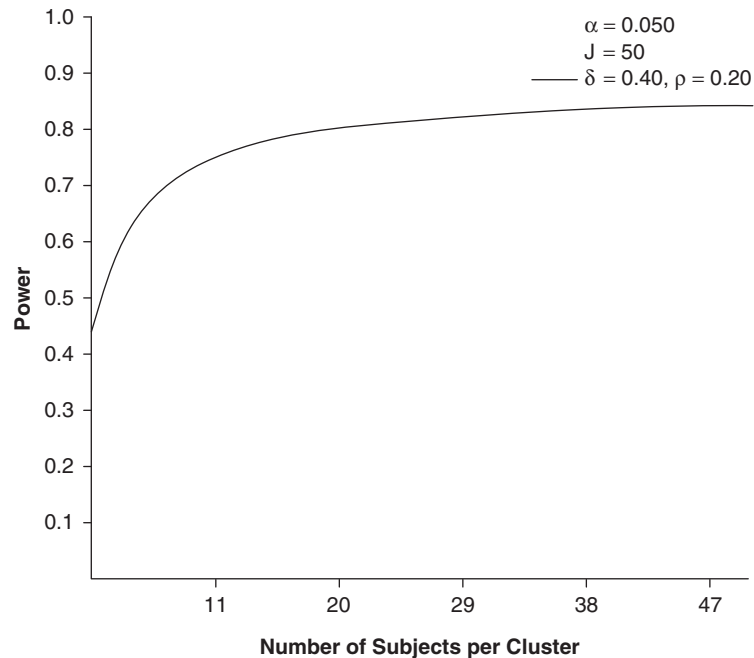


Figure 2.3 The Relationship Between Cluster Size and Power to Detect an Effect Size of .40, With 50 Clusters Total, an ICC of .20, and $\alpha = .05$ (graph generated using optimal design software)

Figure 2.4 depicts the relationship between the number of clusters and the power to detect an effect size of .40 at $\alpha = .05$ for a study with 15 participants per cluster and an ICC of .20. As that graph shows, a power of .80 to detect this effect size is only achieved when the total number of clusters is above 50, and it requires 82 clusters for .95 power. In many research contexts, collecting data from so many clusters may be impractical and other techniques for attaining adequate power must be employed.

Optimizing Power in a Multilevel Design

The techniques for maximizing statistical power in single-level analyses also apply, with appropriate adaptations, to multilevel analyses. Power can be increased by relaxing the alpha level or increasing the sample size (in this case, mainly the number of clusters). Also, adding covariates to the analysis is an effective way to increase power. In multilevel analysis, covariates measured at either the participant level or the cluster level (or both) can be used. Cluster-level covariates are often easier to obtain because each individual participant need not be measured and may be as helpful for increasing power as participant-level covariates (Bloom, 2005; Murray & Blitstein, 2003). As in single-level analysis, one of the best covariates, when available, is the pretest score on the same measure as the outcome variable or

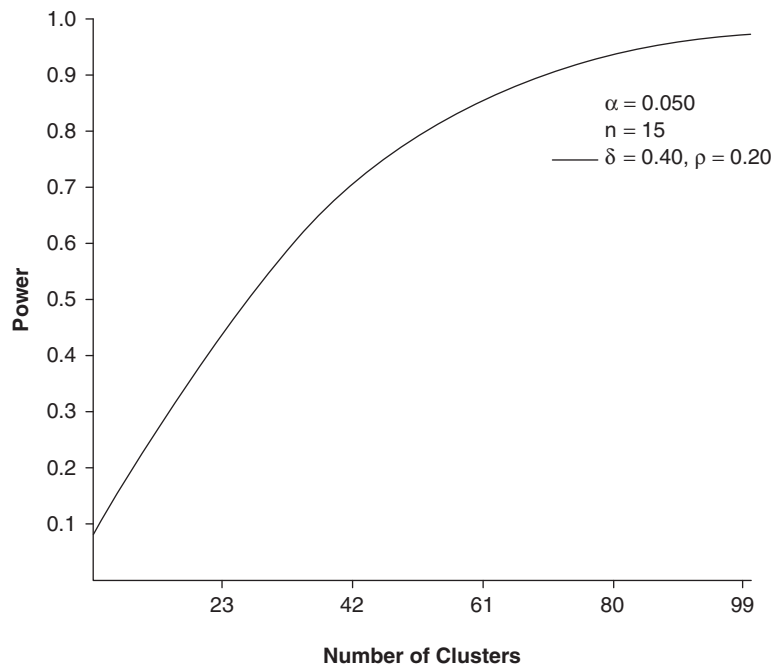


Figure 2.4 The Relationship Between Number of Clusters and Power to Detect an Effect Size of .40, With 15 Participants per Cluster, an ICC of .20, and $\alpha = .05$ (graph generated using optimal design software)

a closely related one. Including a pretest covariate can reduce the number of clusters required to achieve adequate power anywhere from one half to one tenth and cluster-level pretest scores (the mean for each cluster) may be just as useful as participant-level pretest scores (Bloom, Richburg-Hayes, & Black, 2005).

Figure 2.5 illustrates the change in power associated with adding a cluster-level covariate that accounts for varying proportions of the between-cluster variance on the outcome variable. Without a covariate, 52 clusters (26 each in the treatment and control groups) with 15 participants per cluster and an ICC of .20 are required to detect an effect size of .40 at $\alpha = .05$ with .80 power. With the addition of a cluster-level covariate that accounts for 66% of the between-cluster variance (i.e., correlates about .81), the same power is attained with half as many clusters (26 total). Accounting for that proportion of between-cluster variance would require a strong covariate (or set of covariates), but not so strong as to be unrealistic for many research situations.

Planning a Multilevel Study With Adequate Power

Estimating the power of a multilevel study requires taking into account the minimum meaningful effect size that the researcher would like to detect, the alpha level for the statistical test, the number of clusters, the number of participants within

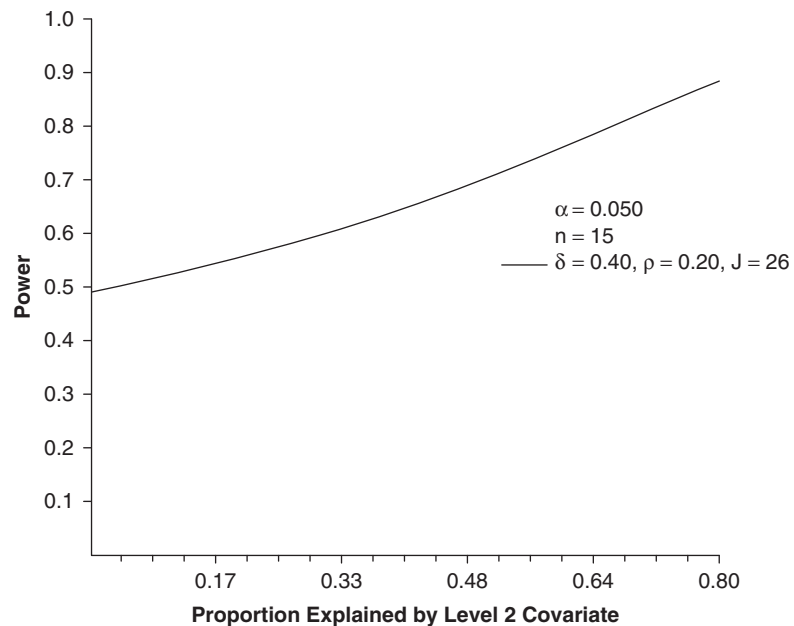


Figure 2.5 Power for Detecting an Effect Size of .40, With 26 Clusters, 15 Participants per Cluster, ICC of .20, and $\alpha = .05$, as Influenced by the Addition of a Cluster-Level Covariate of Various Strengths (graph generated using optimal design software)

each cluster, the ICC associated with those clusters, and any covariates or blocking factors involved in the design. Given all these considerations, it is not surprising that computing power estimates is rather complicated (see Raudenbush, 1997; Snijders & Bosker, 1993, for examples of computational techniques). Fortunately, there is software available that facilitates these computations. One of the best documented and easiest to use is Optimal Design, based on the calculations outlined in Raudenbush and Liu (2000) (available without cost at the time this chapter was written at http://sitemaker.umich.edu/group-based/optimal_design_software). Optimal Design was used to generate the graphs in Figures 2.2, 2.3, 2.4, and 2.5. Power Analysis in Two-Level designs (PINT), developed by Snijders and his colleagues and using the formulas derived in Snijders and Bosker (1993), is another package that provides similar power calculations, but is currently more limited in the research designs that it can accommodate (PINT is available at the time this chapter was written at <http://stat.gamma.rug.nl/snijders>).

Design Strategy to Enhance Power

Perhaps the most important point to be gleaned from the above discussion about statistical power is that nearly all the factors that influence it can be manipulated to

increase power. A research design that is sensitive to intervention effects, therefore, is achieved through the integration of decisions about all these factors in a way that is appropriate and practical for the particular research circumstances. This requires awareness of statistical power issues during the planning phase of a study, incorporation of procedures to enhance power in the design, and an analysis and interpretation of study results that reflects statistical power considerations.

The general strategy for optimizing power in intervention research necessarily begins with a decision about the minimum effect size that the research should be designed to detect reliably (Bloom, 1995). This minimum detectable effect should be set as a threshold value such that below that level, intervention effects are considered too small to be important, but above that level, they are potentially meaningful and thus should be detected by the research. It is at this point that the researcher must consider the various issues related to the effect sizes, such as what treatment versus control contrast will be represented in that effect size. This requires decisions about the “dosage” for the intervention, the nature of the control group (no treatment, placebo, service as usual, and so on), and the character of the dependent variable(s) (e.g., psychometric vs. edumetric).

Given decisions on these points, the researcher must then decide what numerical value of the effect size under the planned research circumstances represents a meaningful minimum to be detected. This usually involves a complex judgment regarding the practical meaning of effects within the particular intervention context. The next section provides some suggestions for framing this issue. For now, suppose that a threshold value has been set: Say that $ES = .20$ is judged the smallest effect size that the research should reliably detect. The next question is how reliably the researcher wishes to be able to detect that value—that is, what level of statistical power is desired. If the desired power is .80, for instance, statistically significant results would be found 80% of the time an effect of .20 was actually present in the populations sampled for the research, and null results would occur 20% of the time despite the population effect. If greater reliability is desired, a higher level of power must be set. Setting the desired power level, of course, is equivalent to setting the beta level for risk of Type II error. Alpha level for Type I error should also be set at this time, using some rational approach to weighing the risks of Type I versus Type II error, as discussed earlier.

With a threshold effect size value and a desired power level in hand, the researcher is ready to address the question of how to actually attain that power level in the research design. At this juncture it is wise to consider what variance control statistics might be used. These can generally be applied at low cost and with only a little extra effort to collect data on appropriate covariate variables or implement blocking. Using the formulas and discussion provided above in the subsection on the statistical test, the researcher can estimate the operative effect size with a variance control design and determine how much larger it will be than the original threshold value. With an ANCOVA design using the pretest as a covariate, for instance, the pretest-posttest correlation might be expected to be at least .80, increasing the operative effect size from the original .20 to a value of .33 (see Table 2.3). Analogous assessments of covariates can be made for multilevel designs by using appropriate statistical power software.

With an operative effect size and a desired power level now established, the researcher is ready to turn to the question of the size of the sample in each experimental group. This is simply a matter of looking up the appropriate value using a statistical power chart or computer program. If the result is a sample size the researcher can achieve, then all is well.

If the required sample size is larger than can be attained, however, it is back to the drawing board for the researcher. The options at this point are limited. First, of course, the researcher may revisit previous decisions and further tune the design—for example, enhancing the treatment versus control contrast, improving the sensitivity of the dependent measure, or applying a stronger variance control design. If this is not possible or not sufficient, all that remains is the possibility of relaxing one or more of the parameters of the study. Alpha or beta levels, or both, might be relaxed, for instance. Because this increases the risk of a false statistical conclusion, and because alpha levels particularly are governed by strong conventions, this must obviously be done with caution. Alternatively, the threshold effect size that the research can reliably detect may be increased. This amounts to reducing the likelihood that effects already assumed to be potentially meaningful will be detected.

Despite best efforts, the researcher may have to proceed with an underpowered design. Such a design may be useful for detecting relatively large effects but may have little chance of detecting smaller, but still meaningful, effects. Under these circumstances, the researcher should take responsibility for communicating the limitations of the research along with its results. To do otherwise encourages misinterpretation of statistically null results as findings of “no effect” when there may be a reasonable probability of an actual effect that the research was simply incapable of detecting.

As is apparent in the above discussion, designing research sensitive to intervention effects depends heavily on an advance specification of the magnitude of statistical effect that represents the threshold for what is important or meaningful in the intervention context. In the next section, we discuss some of the ways in which researchers can approach this judgment.

What Effect Size Is Worth Detecting?

Various frameworks can be constructed to support reasonable judgment about the minimal effect size that an intervention study should be designed to detect. That judgment, in turn, will permit the researcher to consider statistical power in a systematic manner during the design phase of the research. Also, given a framework for judgment about effect size, the researcher can more readily interpret the statistical results of intervention research after it is completed. Below, we review three frameworks for judging effect size: the actuarial approach, the statistical translation approach, and the criterion group contrast approach.

The Actuarial Approach

If enough research exists similar to that of interest, the researcher can use the results of those other studies to create an actuarial base for effect sizes. The distribution of

such effect size estimates can then be used as a basis for judging the likelihood that the research being planned will produce effects of a specified size. For example, a study could reliably detect 80% of the likely effects if it is designed to have sufficient power for the effect size at the 20th percentile of the distribution of effect sizes found in similar studies.

Other than the problem of finding sufficient research literature to draw on, the major difficulty with the actuarial approach is the need to extract effect size estimates from studies that typically do not report their results in those terms. This, however, is exactly the problem faced in meta-analysis when a researcher attempts to obtain effect size estimates for each of a defined set of studies and do higher-order analysis on them. Books and articles on meta-analysis techniques contain detailed information about how to estimate effect sizes from the statistics provided in study reports (see, e.g., Lipsey & Wilson, 2000).

A researcher can obtain a very general picture of the range and magnitude of effect size estimates in intervention research by examining any meta-analyses that have been conducted on similar interventions. Lipsey and Wilson (1993) reported the distribution of effect sizes from more than 300 meta-analyses of research on psychological, behavioral, and educational research. That distribution had a median effect size of .44, with the 20th percentile at .24 and the 80th percentile at .68. These values might be compared with the rule of thumb for effect size suggested by Cohen (1977, 1988), who reported that across a wide range of social science research, $ES = .20$ could be judged as a “small” effect, $.50$ as “medium,” and $.80$ as “large.”

The Statistical Translation Approach

Expressing effect sizes in standard deviation units has the advantage of staying close to the terms used in statistical significance testing and, thus, facilitating statistical power analysis. However, that formulation has the disadvantage that in many intervention domains there is little basis for intuition about the practical meaning of a standard deviation's worth of difference between experimental groups. One approach to this situation is to translate the effect size index from standard deviation units to some alternate form that is easier to assess.

Perhaps the easiest translation is simply to express the effect size in the units of the dependent measure of interest. The ES index, recall, is the difference between the means of the treatment and control groups divided by the pooled standard deviation. Previous research, norms for standardized tests, or pilot research is often capable of providing a reasonable value for the relevant standard deviation. With that value in hand, the researcher can convert to the metric of the specific variable any level of ES he or she is considering. For example, if the dependent variable is a standardized reading achievement test for which the norms indicate a standard deviation of 15 points, the researcher can think of $ES = .50$ as 7.5 points on that test. In context, it may be easier to judge the practical magnitude of 7.5 points on a familiar test than $.50$ standard deviations.

Sometimes, what we want to know about the magnitude of an effect is best expressed in terms of the proportion of people who attained a given level of benefit as a result of intervention. One attractive way to depict effect size, therefore,

70 APPROACHES TO APPLIED RESEARCH

is in terms of the proportion of the treatment group, in comparison to the control group, elevated over some “success” threshold by the intervention. This requires, of course, that the researcher be able to set some reasonable criterion for success on the dependent variable, but even a relatively arbitrary threshold can be used to illustrate the magnitude of the difference between treatment and control groups.

One general approach to expressing effect size in success rate terms is to set the mean of the control group distribution as the success threshold value. With symmetrical normal distributions, 50% of the control group will be below that point and 50% will be above. These proportions can be compared with those of the treatment group distribution below and above the same point for any given difference between the two distributions in standard deviation units. Figure 2.6 depicts the relationship for an effect size of $ES = .50$. In this case, 70% of the treatment group is above the mean of the control group, or, in failure rate terms, only 30% of the treated group is below the control group mean. There are various ways to construct indices of the overlap between distributions to represent effect size. This particular one corresponds to Cohen’s (1977, p. 31) $U3$ measure.

A variation on the percentage overlap index has been offered by Rosenthal and Rubin (1982), who used it to construct something that they call a “binominal effect size display” (BESD). They suggest that the success threshold be presumed to be at the grand median for the conjoint control and treatment distribution (line M in Figure 2.6). Though use of the grand median as a success threshold is somewhat arbitrary, it confers a particular advantage on the BESD. With normal distributions, the difference between the “success” proportions of the treatment and control groups has a simple relationship to the effect size expressed in correlational terms. In particular, when we express effect size as a correlation (r), the value of that correlation corresponds to the difference between the proportions of the respective distributions that are above the grand median success threshold. Effect size in standard deviation units can easily be converted into the equivalent correlation using the following formula:

$$r = \frac{ES}{\sqrt{ES^2 + 4}}$$

For example, if the correlation between the independent variable and the dependent variable is .24, then the difference between the success proportions of the groups is .24, evenly divided around the .50 point, that is, $.50 \pm .12$, or 38% success in the control group, 62% in the treatment group. More generally, the distribution with the lower mean will have $.50 - (r/2)$ of its cases above the grand median success threshold, and the distribution with the greater mean will have $.50 + (r/2)$ of its cases above that threshold. For convenience, Table 2.4 presents the BESD terms for a range of ES and r values as well as Cohen’s $U3$ index described above.

The most striking thing about the BESD and the $U3$ representations of the effect size is the different impression that they give of the potential practical significance of a given effect from that of the standard deviation expression. For

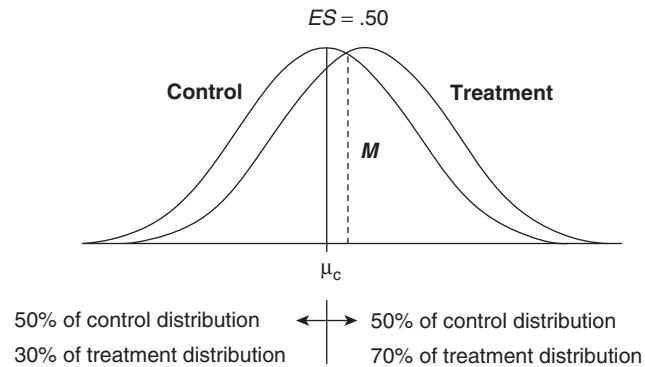


Figure 2.6 Depiction of the Percentage of the Treatment Distribution Above the Success Threshold Set at the Mean of the Control Distribution

Table 2.4 Effect Size Equivalents for ES , r , $U3$, and $BESD$

| ES | r | $U3$: % of T Above X_c | $BESD$ C Versus T | | $BESD$ C Versus T Differential |
|------|-----|--------------------------------|-------------------|-----|-----------------------------------|
| | | | Success Rates | | |
| .10 | .05 | 54 | .47 | .52 | .05 |
| .20 | .10 | 58 | .45 | .55 | .10 |
| .30 | .15 | 62 | .42 | .57 | .15 |
| .40 | .20 | 66 | .40 | .60 | .20 |
| .50 | .24 | 69 | .38 | .62 | .24 |
| .60 | .29 | 73 | .35 | .64 | .29 |
| .70 | .33 | 76 | .33 | .66 | .33 |
| .80 | .37 | 79 | .31 | .68 | .37 |
| .90 | .41 | 82 | .29 | .70 | .41 |
| 1.00 | .45 | 84 | .27 | .72 | .45 |

example, an effect size of one fifth of a standard deviation ($ES = .20$) corresponds to a $BESD$ success rate differential of .10—that is, 10 percentage points between the treatment and control group success rates (55% vs. 45%). A success increase of 10 percentage points on a control group baseline of 45% represents a 22% improvement in the success rate ($10/45$). Viewed in these terms, the same intervention effect that may appear rather trivial in standard deviation units now looks potentially meaningful.

The Criterion Contrast Approach

Although actuarial and statistical translation approaches to assessing effect size may be useful for many purposes, they are somewhat removed from the specific context of any given intervention study. Often, the best answer to the question of what effect size has practical significance is one that is closely tied to the particular problems, populations, and measures relevant to the intervention under investigation. For example, if we could identify and measure a naturally occurring effect in the intervention context whose practical significance was easily recognized, it could be used as a criterion value or benchmark against which any expected or obtained intervention effect could be compared. What is required in the criterion group contrast approach is that some such comparison be identified and represented as a statistical effect size on the dependent measure relevant to the intervention research.

The criterion group contrast approach is best explained by an example. Consider a community mental health center in which prospective patients receive a routine diagnostic intake interview and are sorted into those judged to need, say, inpatient therapy versus outpatient therapy. This practice embodies a distinction between more serious and less serious cases and the “size” of the difference between the severity of the symptoms for these two groups that would be well understood at the practical level by those involved in community mental health settings. If we administer a functional status measure that is of interest as an outcome variable for both these groups, we could represent the difference between them as an effect size—that is, the difference between their means on that measure divided by the pooled standard deviations. Though this effect size does not represent the effect of intervention, we can nonetheless think of it in comparison with an intervention effect. That is, how successful would we judge a treatment to be that, when applied to clients as severe as the inpatient group, left them with scores similar to those of the outpatient group? Such an effect may well be judged to be of practical significance and would have recognized meaning in the treatment context. Real or anticipated intervention effects can thus be compared with this criterion contrast value as a way of judging their practical significance.

Reasonable criterion comparisons are often surprisingly easy to find in applied settings. All one needs to create a criterion contrast are, first, two groups whose difference on the variable of interest is easily recognized and, second, the result of measurement on that variable. It is also desirable to use groups that resemble, as much as possible, those samples likely to be used in any actual intervention research. Some of the possibilities for criterion contrasts that frequently occur in practical settings include the following:

- Eligible versus ineligible applicants for service where eligibility is determined primarily on the basis of judged need or severity. For example, a contrast on economic status might compare those who do not qualify for food stamps with those who do.
- Sorting of intervention recipients into different service or diagnostic categories based on the severity of the problems to be treated. For example, a contrast

on literacy might compare those adult education students enrolled in remedial reading classes with those enrolled in other kinds of classes.

- Categories of termination status after intervention. For example, a contrast on functional status measures might compare those patients judged by physical therapists to have had successful outcomes with those judged to have had unsuccessful outcomes.
- Comparison of “normal” individuals with those who have the target problem. For example, a contrast on delinquent behavior could compare the frequency of self-reported delinquency for a sample of males arrested by the police with that of similar-age males from a general high school sample.
- Maturation differences and/or those occurring with usual service. For example, a contrast on mathematics achievement might compare the achievement test scores of third graders with those of fifth graders.

Conclusion

Attaining adequate statistical power in intervention research is not an easy matter. The basic dilemma is that high power requires a large effect size, a large sample size, or both. Despite their potential practical significance, however, the interventions of interest all too often produce modest statistical effects, and the samples on which they can be studied are often of limited size. Intervention researchers need to learn to live responsibly with this problem. The most important elements of a coping strategy are recognizing the predicament and attempting to overcome it in every possible way during the design phase of a study. The keys to designing sensitive intervention research are an understanding of the factors that influence statistical power and the adroit application of that understanding to the planning and implementation of each study undertaken. As an aid to recall and application, Table 2.5 lists the factors discussed in this chapter that play a role in the statistical power of experimental research along with some others of an analogous sort.

Table 2.5 Factors That Work to Increase Statistical Power in Treatment Effectiveness Research

Independent variable

- Strong treatment, high dosage in the treatment condition
- Untreated or low-dosage control condition for high contrast with treatment
- Treatment integrity; uniform application of treatment to recipients
- Control group integrity; uniform control conditions for recipients

Study participants

- Large sample size (or number of clusters in the case of multilevel research) in each experimental condition

(Continued)

74 APPROACHES TO APPLIED RESEARCH

Table 2.5 (Continued)

Deploying limited participants into few rather than many experimental groups
 Little initial heterogeneity on the dependent variable
 Measurement or variance control of participant heterogeneity
 Differential participant response accounted for statistically (interactions)

Dependent variables

Validity for measuring characteristic expected to change
 Validity, sensitivity for change on characteristic measured
 Fine-grained units of measurement rather than coarse or categorical
 No floor or ceiling effects in the range of expected response
 Mastery or criterion-oriented rather than individual differences measures
 Inherent reliability in measure, unresponsiveness to irrelevant factors
 Consistency in measurement procedures
 Aggregation of unreliable measures
 Timing of measurement to coincide with peak response to treatment

Statistical analysis

Larger alpha for significance testing
 Significance tests for graduated scores, not ordinal or categorical
 Statistical variance control; blocking, ANCOVA, interactions

Discussion Questions

1. In your area of research, which type of error (Type I or Type II) typically carries more serious consequences? Why?
2. In your field, would it ever be sensible to perform a one-tailed significance test? Why or why not?
3. In your field, what are some typical constructs that would be of interest as outcomes, and how are those constructs usually measured? What are the pros and cons of these measures in terms of validity for measuring change, reliability, and discrimination of individual differences?
4. In your research, what are some extraneous factors that are likely to be correlated with your dependent variables? Which of these are measurable so that they might be included as covariates in a statistical analysis?
5. What are some ways that you might measure implementation of an intervention in your field of research? Is it likely that interventions in your field are delivered uniformly to all participants?
6. Is the use of “no treatment” control groups (groups that receive no form of intervention) typically possible in your field? Why or why not?

7. In your field, are interventions typically delivered to individual participants, or to groups of participants such as classrooms, neighborhoods, etc.? If interventions are delivered to groups, do researchers normally use analytical techniques that take this into account?

8. If you were designing a study in which an intervention was to be delivered to groups (clusters) of participants, would you be better off, in terms of statistical power, collecting data on a large number of individuals within each cluster or on a smaller number of individuals in a larger number of clusters?

9. Imagine you conduct a study testing an intervention that is designed to increase the intelligence of children. You have access to a very large number of children and, thus, have adequate power to detect an effect size of .03. At the end of the intervention, the average IQ score of children in your control group is 100.0, and the average IQ score of children in your intervention group is 100.5. This difference in IQ scores is statistically significant. What do you conclude from your study?

Exercises

1. Look up four or five recent studies with treatment/control comparisons in your area of research and calculate the effect sizes they report. What is the average effect size, and what is the range of effect sizes? If you were designing a similar study, what is the minimum effect size that you would consider meaningful to detect?

2. Using the power chart in Figure 2.1, determine the power to detect an effect size of .70 with 20 participants per group, given a two-tailed α of .05. How many participants per group would you need to attain .90 power to detect the same effect size?

3. You are designing a study examining gains on a standardized test of academic achievement and your research leads you to believe that you can expect an effect size of .30 (assume the intervention group mean will be 105, the control group mean 100, and the shared standard deviation 15). Unfortunately, constraints on your resources require a design that is able to detect a minimum effect size of .60. If you were to add a covariate to your model to increase power, how strongly must that covariate be correlated with academic achievement to give you adequate power, given your design constraints?

References

- Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 19(5), 547–556.
- Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York: Russell Sage Foundation.

76 APPROACHES TO APPLIED RESEARCH

- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions* (MDRC Working Papers on Research Methodology). New York: MDRC.
- Brown, G. W. (1983). Errors, Type I and II. *American Journal of Disorders in Childhood, 137*, 586–591.
- Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist, 29*, 512–518.
- Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology, 36*, 517–526.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Hedges, L. V., & Hedberg, E. C. (2006). *Intraclass correlation values for planning group randomized trials in education* (Institution for Policy Research Working Paper). Evanston, IL: Northwestern University.
- Hox, J. (2002) *Multilevel Analysis: Techniques and Applications*. Hillsdale, NJ: Lawrence Erlbaum.
- Kraemer, H. C., & Thieman, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.
- Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Murphy, K. R., & Myers, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review, 27*, 79–103.
- Nagel, S. S., & Neef, M. (1977). Determining an optimum level of statistical significance. In M. Guttentag & S. Saar (Eds.), *Evaluation studies review annual* (Vol. 2, pp. 146–158). Beverly Hills, CA: Sage.
- Rasbash, J., Steele, F., Browne, W. J., & Prosser, B. (2004). *A user's guide to MLwiN* (Version 2.0). London: Institute of Education.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173–185.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: SSI.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166–169.
- Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation Review, 8*, 573–582.
- Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237–259.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Sage.