NCSER 2010-3006

U. S. DEPARTMENT OF EDUCATION

Statistical Power Analysis in Education Research

ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Statistical Power Analysis in Education Research

APRIL 2010

Larry V. Hedges
Christopher Rhoads
Northwestern University

Abstract

This paper provides a guide to calculating statistical power for the complex multilevel designs that are used in most field studies in education research. For multilevel evaluation studies in the field of education, it is important to account for the impact of clustering on the standard errors of estimates of treatment effects. Using ideas from survey research, the paper explains how sample design induces random variation in the quantities observed in a randomized experiment, and how this random variation relates to statistical power. The manner in which statistical power depends upon the values of intraclass correlations, sample sizes at the various levels, the standardized average treatment effect (effect size), the multiple correlation between covariates and the outcome at different levels, and the heterogeneity of treatment effects across sampling units is illustrated. Both hierarchical and randomized block designs are considered. The paper demonstrates that statistical power in complex designs involving clustered sampling can be computed simply from standard power tables using the idea of operational effect sizes: effect sizes multiplied by a design effect that depends on features of the complex experimental design. These concepts are applied to provide methods for computing power for each of the research designs most frequently used in education research.

NCSER 2010-3006
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Special Education Research, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer

The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Optimal Solutions Group, LLC to develop a guide for calculating statistical power for complex multilevel designs that are used in most field studies in education. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

April 2010

Alternate Formats

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

There are two authors for this report with whom IES contracted to develop the discussion of the issues presented. Dr. Larry Hedges and Dr. Christopher Rhoads are both employees of Northwestern University. The authors do not have financial interests that could be affected by the content in this report.

Contents

List of Tables

Chapter 1: Introduction

Experimental evaluation seeks to make possible valid inferences about the effects of a treatment, or intervention, in question. However, a research study can make an invalid inference by concluding that a treatment has an effect when it does not: This mistaken conclusion is called a Type I error in statistical decision theory. Statistical significance testing is designed to control the chance of making Type I errors. A second way that a research study can make an invalid inference is by failing to detect that a treatment has an effect when the true treatment effect is nonzero. This is called a Type II error in statistical decision theory. Statistical power is defined as the probability that a research study avoids a Type II error by correctly rejecting a null hypothesis of zero treatment effect. Low statistical power increases the probability of obtaining a Type II error and is a major threat to the statistical conclusion validity of educational research studies (Shadish, Cook, and Campbell 2002).

Statistical power analysis is a method of determining the probability that a proposed research design will detect the anticipated effects of a treatment. It helps the researcher determine whether a study design should be modified so that it will have adequate power for detecting effects. The purpose of this paper is to provide an introduction to the computation of statistical power for education field studies and to discuss research design parameters that directly impact statistical power. Most field studies in education involve complex designs, typically involving clustered sampling of students within classes or schools and assignment to treatments by groups (such as classrooms or schools). As will be expanded upon later in the paper, clustering directly impacts statistical power because the relationships among units within clusters usually implies a greater similarity of outcomes for units within a cluster than for units coming from different clusters. For example, if there were two third-grade classrooms in a school, one would expect greater similarities among students within each of the classrooms than among students from different classrooms. This concept, represented statistically by the intraclass correlation coefficient (ICC), plays a crucial role in the power analysis of designs with clustering (Donner and Klar 2000; Murray 1998; Shadish, Cook, and Campbell 2002).

This paper is intended for education researchers who are familiar with basic statistics concepts but do not consider themselves to be experts in statistics. Often, education researchers have training in statistical power analysis that is limited to studies that have relatively simple designs (e.g., one level of sampling and individual randomization). This paper provides a guide for calculating statistical power for more complicated multilevel designs that are used in most field studies in education.

This paper begins with an explanation of sampling theory applied to education research and shows that sample design in education research can be understood in terms of ideas from survey research. This conceptual transferability is possible because, similar to survey data and designs, the data considered in many—if not most—educational applications are not obtained from a simple random sample, but rather from a more complex sample design. As a result, the analyses of the research and assumptions about the research designs must be addressed accordingly. Understanding how sample design induces random variation in the quantities observed in a randomized experiment is a necessary precursor to understanding statistical power.

Next, the topics of hierarchical designs and randomized-block designs—the two major classes of experimental designs that are predominantly used in education research—are explicated to explore the concept of multiple levels of analysis and how to address them in education research. The concept of statistical power is then introduced. This paper contains demonstrations that statistical power in complex designs involving clustered sampling can be computed simply using the idea of operational effect sizes—effect sizes multiplied by a design effect that depends on features of the complex experimental design. Finally, these concepts are applied to provide methods for computing power for each of the 10 research designs most frequently used in education research.

This paper has five appendices. Appendix A provides formulas for computing design effects in multilevel hierarchical designs. Appendix B provides formulas for computing design effects in multilevel randomized-block designs. Appendix C details methods for computing power in three-level randomized-block designs. Appendix D describes the multilevel models on which power computations are based. Finally, Appendix E provides a glossary of technical terms and the page numbers on which these terms are used. This paper also contains two tables that can be used for determining the power of the test for treatment effects in commonly used education research designs.

Readers should remember that assumptions and values used in the demonstrations throughout the report are intended for illustration only and are not intended to suggest reference values to use when planning other research studies.

Chapter 2: Sampling and Sample Design

Statistical analyses are tools for drawing inferences in the presence of random fluctuations or uncertainty ("randomness" and "uncertainty" are used synonymously in the field of research, but for the sake of simplicity, "randomness" will be used throughout this paper). The conceptual models on which statistical analysis methods are based depend on the idea that the randomness comes from sampling. Hence, an understanding of how sample design induces random variation in the quantities observed in a randomized experiment is a prerequisite for understanding statistical power. This crucial point is often misunderstood.

Sample Designs, Sample Surveys, and Hierarchical Structures

The sampling design is the process by which individuals from the population are selected for the sample. Most statistical and education research—and most statistical theory—begins with the assumption that the data being considered come from a simple random sample. Unfortunately, the data considered in many, if not most, educational applications are not obtained from a simple random sample but rather from a more complex sample design. This section reviews two crucial concepts from sample design—randomized-block design and hierarchical design—and applies them to designs for education research.

A population of potential observations can often be identified as having some structure that makes it non-uniform in obvious ways. In survey research, this structure is often geographic (e.g., states; smaller geographic divisions, such as census tracts; or still smaller geographic divisions, such as neighborhood blocks), but it could also be demographic (e.g., groups defined by age, gender, or race). In education studies of children in schools, the relevant population structure is often defined by the hierarchical organization of the educational system (i.e., children are nested within classrooms, classrooms are nested within schools, and schools are nested within school districts).

The fact that populations are structured (and the same population may be structured in several different ways) does not necessarily mean that simple random sampling is impossible, nor does it affect the properties of statistics computed from simple random samples of these populations. Population structure, in fact, can provide opportunities to collect samples in ways that have practical advantages. For example, a simple random sample of fourth-grade students in a state would begin with a list of all fourth-graders in the state and then use a random device to pick the required number of students from the list. This sampling would typically result in a sample with very few students in any one classroom or school and a relatively large number of schools. Such a sample would be difficult and expensive to collect because it would involve obtaining data from many different classrooms and schools. Additionally, it may be impossible to obtain a list of all fourth-graders in the state but easy enough to obtain a list of all schools in the state. In this case, one may obtain a sample of fourth-grade students in two stages. In the first stage, a simple random sample of schools is selected. In the second stage, a sample of fourth-graders is collected from within each school selected in the first stage of sampling. Samples obtained in this manner are called hierarchical or clustered samples. Thus, clustered sampling designs provide a mechanism for obtaining a random sample from the desired population when a simple random sampling design is either impossible or not cost effective.

Stratification and Clustered Sampling

Sample survey designs almost always use one of two methods to simplify data collection. Which of the two methods is used has important implications for statistical analysis. One method is stratification, which is the division of the population into subsets (i.e., strata) that do not overlap (such as gender or achievement level) (Shadish, Cook, and Campbell 2002). The sample is then drawn so that some individuals are selected within each stratum. The second method used to simplify data collection is clustered sampling. Clustered sampling also begins with the division of the population into non-overlapping subsets (now called clusters), but in the case of clustered sampling, individuals are not strategically or intentionally sampled from each of the clusters; instead, only individuals from a sample of the clusters are selected into the sample. Operationally, a cluster sample is drawn in two stages, which is why clustered sampling is often called a two-stage or two-level sampling. In the first stage, a simple random sample of clusters, such as schools, is selected, and then a simple random sample of individuals, such as students within these schools, is drawn within each of the clusters that were selected at the first stage. The distinction between stratified sampling and clustered sampling lies in whether some individuals are included in the sample from *every one* of the subdivisions of the population (such as gender and achievement level). In other words, if every one of the clusters in the population is *intentionally* included in the sample, the cluster sample becomes a stratified sample. Therefore, stratified samples include individuals from each of the subdivisions, and clustered samples do not.

Clustered sampling involving clusters at more than one level of the population is also possible, and this practice is widely used in survey sampling. For example, one might choose first to sample school districts (clusters at one level), then sample schools (subclusters at a lower level), and then sample children within schools at the third level. Alternatively, one might sample schools (clusters at one level), then sample classrooms (subclusters at a lower level), and then sample children within classrooms at the third level. Such sampling designs are called three-stage samples or three-level sampling designs. In the case of two-level or three-level sampling designs, whether the subdivisions of the population are clusters or strata depends upon whether *all* of the subdivisions in the population are included in the sample. This determination, in turn, depends on the definition of the population of interest.

Population of Interest

The concept of population of interest is more conceptually slippery in intervention studies than in sample surveys. In sample surveys, the population of interest is typically a static population at some moment in time, such as high school graduates, or students needing a specific service, such as a behavioral intervention. In such cases, it is fairly easy to determine what all the population subdivisions might be and whether they are all included in the sample. In intervention studies, however, the population may or may not be a static population. For instance, a population of students could be identified as requiring intensive intervention, based on a score on a screening measure. Once these students had received the intervention, however, it is possible that some of them would be designated as no longer needing the intervention because they had reached certain benchmarks. Thus, the population in this case is not necessarily static.

In some cases, the population of interest may be fixed in both time and place. The most important example is in studies where the primary question may be, "Did the treatment produce an effect on the students studied in the classrooms and schools that were part of the study?" For instance, did an intervention work well with displaced survivors of Hurricane Katrina in New Orleans who are eligible for counseling services in the elementary grades? In such cases, it is reasonable to consider the particular schools and classrooms in the study to define the population of interest. Only the sampling of students into schools and classrooms is a source of sampling randomness. Inferences are, therefore, restricted to—or conditional on—the particular schools and classrooms in the study. The inference model associated with this population definition is often called the conditional inference model (Hedges and Vevea 1998). In other cases, the population of interest may not be fixed in either time or place. The most obvious examples are effectiveness studies in which the object is to determine whether the intervention would produce effects in a wider (perhaps national) population of schools and classrooms. For instance, will the intervention work with all survivors of natural disasters who are in need of counseling services in the elementary grades? In such cases, it is natural to consider the particular schools and classrooms in which the intervention takes place as only a sample of schools and classrooms from the larger population to which one might generalize. The inference model associated with this population definition is often called the unconditional inference model (Hedges and Vevea 1998).

There are further subtleties in these population definitions when considering population subdivisions, such as classrooms within schools (or if districts are assigned to treatments, schools within districts). Suppose that a particular school has only three fourth-grade classrooms. One might take the position that all of the fourth-grade children in the school at the present time are in one of these classrooms. Consequently, the three classrooms represent three strata of the school population. Alternatively, one might take the position that these three classrooms are just the three classrooms that happen to be in the school at the present time. Following this train of thought, one could conclude that there will be other classrooms in the school in the future and the three that happen to be there at the present time are a sample from a population of possible classrooms within that school.

This latter argument may seem strange until one considers the analogous argument about students. Although all of the students within the classroom may be sampled, it would seem odd to say that the entire population of students who could have been in that classroom has been sampled. The more natural population definition seems to require that the students be considered a sample from a population of students who *could* have been sampled into that classroom. Similarly, if one imagines that the test scores of students in a particular classroom are influenced by the teacher they happen to have, then it is odd to think that the particular classrooms in a school at any one time—and the teachers that happen to be assigned to them—constitute a population of interest. A more natural population of interest seems to be one where the present teachers (and therefore classrooms) are a sample of potential teachers (and therefore classrooms).

This paper focuses on the unconditional inference model, in which the observed set of population subdivisions is considered a sample of the larger population. Thus, schools and classrooms will be considered clusters in multistage clustered samples. The statistical analyses associated with

this inference model would naturally include schools and classrooms as random effects if they were considered clusters in the sampling design.

Implications for Statistical Analysis

If the details of population definition, such as the distinction between strata and clusters (or the conditional versus unconditional inference models), were merely a matter of terminology, it would be of little scientific consequence. However, the distinction between clusters and strata is of major consequence to the sampling distribution of statistics computed from a sample. In order to make this distinction precise, it is necessary to introduce notation for certain population quantities and a specific model for quantifying the randomness about these quantities contained in any sample.

Suppose that the population has a nested structure of individuals within schools (clusters), that the observations within the schools (clusters) are normally distributed about cluster means with a common within-cluster variance σ_W^2, and that the cluster means themselves are normally distributed about the overall population mean μ with between-cluster variance σ_S^2. The total variance of the observations in the population is, therefore,

$$\sigma_T^2 = \sigma_S^2 + \sigma_W^2.$$

The amount of clustering by schools in the population is quantified by the intraclass correlation coefficient (ICC). The ICC describes the extent to which the students within a cluster (e.g., schools) are more alike than those in different clusters (e.g., different schools within the same district).

$$\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2} = \frac{\sigma_S^2}{\sigma_T^2}. \qquad (1)$$

Consider a simple random sample of size mn. A result from elementary statistics is that the variance of the mean of that simple random sample will be

$$\sigma_T^2/mn.$$

Now suppose that instead of a simple random sample, a sample of the same size mn is obtained by first sampling m schools (clusters) and then obtaining a simple random sample of n individuals within each school (cluster). The variance of the mean of this clustered sample would not be σ_T^2/mn but

$$[\sigma_T^2/mn][1 + (n - 1)\rho].$$

The variance of the mean of the clustered sample is bigger by a factor $[1 + (n - 1)\rho]$, which is sometimes called the sample design effect (Kish 1965) or, more descriptively, the variance inflation factor. In this paper, however, the term “design effect” will be used to describe a

somewhat different quantity. We use “design effect” to describe the gain in precision obtained by having sampled more than one individual per cluster.

The fact that sample means from clustered samples have more sampling randomness than those from unclustered (simple random) samples has implications for experimental design. Because treatment effects are defined as mean differences, when entire clusters are assigned to treatments, estimates of treatment effects are less precise than estimates from simple random samples of the same size.

Chapter 3: Experimental Designs

Although many experimental designs have been developed (see Kirk 1995), the two most widely used experimental designs in education research are variants of one of two basic designs: the hierarchical (or nested) design and the (generalized) randomized-block design. These designs both assume a clustered sampling design, but they differ in how random assignment is made to treatments. Because nearly all field studies use a variant of one of these two basic designs, this paper will limit its discussion to these two designs.

The Hierarchical Design

In the hierarchical design, entire clusters (e.g., schools or classrooms) are assigned to treatments (Kirk 1995). Thus, every student in a cluster (a school or classroom) receives the same treatment, but students in different clusters may receive different treatments (e.g., grade-level mathematics tutoring). This design is sometimes called the cluster-randomized design because entire clusters are randomly assigned to treatments. The hierarchical design is perhaps the most widely used experimental design in education; it is desirable in situations where contamination among treatment groups would be possible if more than one treatment were present in the same cluster (e.g., in the same classroom). The hierarchical design minimizes potential problems of contamination between treatments because only one treatment is present in the same cluster (e.g., in the same classroom). In other words, this type of design helps to alleviate contamination because the whole cluster (e.g., the classroom) receives the treatment. The hierarchical design is also desirable in situations where it would be practically difficult to assign the treatment to some students in a cluster but not to others. Some treatments act at the level of the entire cluster (e.g., whole-school interventions such as positive behavior support). In these cases, it would be conceptually impossible to assign different treatments to different individuals within a cluster or to withhold treatment from one group but grant it to another within a cluster.

Every hierarchical design involves at least one stage of clustered sampling (such as sampling schools first and then students within schools). However, the hierarchical design may also involve two or more stages of clustered sampling. For example, if schools are sampled first, then classrooms, then students within classrooms, the result is a three-stage sample with two levels of clustering. In hierarchical designs, random assignment to treatments occurs only to entire clusters at the highest level of clustering. For example, schools would be the unit of random assignment in a three-level design involving schools, classes, and students. If individuals within the same cluster are assigned to different treatments, the design is no longer considered a hierarchical design but instead is referred to as a randomized-block design, as described in the next section.

The Randomized-Block Design

In the randomized-block design, individuals within the same cluster are assigned to different treatments. For example, if students within the same school are assigned to either of two treatments (e.g., an intervention and a control), the design is a generalized randomized-block design (Kirk 1995). The randomized-block design is sometimes called a matched design because individuals assigned to treatments are matched within clusters or blocks. This design is not to be confused with quasi-experimental designs, which are also called matched designs. This paper

uses the term “matched design” to express that the students are not assigned to treatments as individuals but rather as a part of a block or cluster. A matched design is also sometimes called a multisite design because an important application is to multisite trials (such as multicenter medical trials) where the clusters are sites. This design has the advantage that treatment effects are estimated within clusters so that variation among clusters can be estimated separately from variation in treatment effects, and the variation among clusters does not increase the variation in the estimated average treatment effect. Relating this design to education research, one could say that when assignment is made to individuals within schools, school-specific attributes can be estimated separately from school-specific treatment effects.

Every randomized-block design involves at least one stage of clustered sampling (such as sampling schools first and then students within schools before they are assigned to treatments within schools). Like hierarchical designs, randomized-block designs may also have more than one stage of clustering in the sample. For example, schools may be sampled first, then classrooms, and then students within classrooms. Given this three-stage sampling plan with two levels of clustering, there are two ways to assign treatments within clusters, leading to two different experimental designs. One design assigns entire intact classrooms to treatments, so that every individual within the same classroom receives the same treatment. The other design assigns individual students within classrooms to treatments, so that different students within the same classroom may have different treatments. Regardless of whether there are two or three levels of clustering, assignment to treatments must occur *within* the highest level clusters in the sampling design. If it does not (that is, if assignment to treatments occurs at the highest level of clustering), the design becomes a hierarchical design.

Chapter 4: Statistical Power

Research studies evaluating education interventions are typically required to demonstrate that they are designed sufficiently well to provide sound evidence about the effects of the intervention in question. Many factors contribute to a successful intervention design, including feasibility of recruiting and retaining the sample, likelihood of successful implementation of the intervention, and adequate measurement of the outcome. One fundamental characteristic of a research design is whether it will have a good chance to detect the expected effect of the intervention or, alternatively, to detect the smallest effect that is deemed to be educationally meaningful. Statistical power is the probability that a test of the null hypothesis of no treatment effect will successfully reject the null hypothesis when a nonzero average treatment effect exists. In simple research designs that use simple random samples, statistical power (e.g., Cohen 1977) depends on three things:

- the significance level of the test;
- the expected size of the intervention effect (the effect size); and
- the sample size.

In multilevel research designs that involve clustering within schools or classrooms, power also depends on two other factors that are unique in multilevel designs:

- how the sample size is distributed over the levels of the design (the sample size at each level); and
- the extent of the clustering effects (typically measured by one or more intraclass correlation coefficients [ICCs] in hierarchical designs and by "heterogeneity" parameters that quantify the extent to which treatment effects vary across clusters in randomized-block designs).

In designs with clustered samples, many different configurations of sample size at each level can lead to the same total sample size, and not all of these configurations lead to the same statistical power (Konstantopoulos 2008a; Konstantopoulos 2008b; Raudenbush 1997; Raudenbush and Liu 2000; Snijders 2005). For example, in a hierarchical design that assigns schools to treatments, one can achieve a total sample size of $N = 1{,}000$ by assigning to treatments $m = 10$ schools with $n = 100$ students each or $m = 100$ schools of $n = 10$ students each. If there were no clustering effects, both choices would lead to identical statistical power. If there is clustering, these two choices of sample size allocation lead to very different statistical power.

In designs with clustered samples, the extent of clustering may vary, depending on the population studied, age level, subject matter, and outcome measured. The degree of clustering is usually measured by comparing the variation among clusters to the total variation via an ICC such as that described earlier. Such ICCs measure (on a range from zero to one) the extent to which information within clusters is redundant. If the ICC is near zero, there is little clustering and there is little redundant information within clusters. In this case, the statistical power will be close to that of a design that used simple random sampling and the same total sample size. When the ICC is near one, information within clusters is highly redundant, and thus, most of the

variation is between cluster means. In this case, the statistical power will be close to that of a design that used simple random sampling and a total sample size equal to the number of *clusters*. In other words, if the ICC is close to zero, the clusters are quite similar. When the ICC is close to one, the clusters are quite dissimilar, and thus, the variation between cluster means is much higher than the variation within the clusters.

Use of Covariates to Increase Statistical Power

One additional factor that can profoundly influence statistical power—in both simple designs and cluster-randomized designs—is whether covariates are used in the design to increase precision. In multilevel designs, as in single-level designs, the use of covariates can dramatically increase statistical power (Bloom, Richburg-Hayes, and Black 2007; Hedges and Hedberg 2007; Raudenbush, Martinez, and Spybrook 2007). Covariates increase power in multilevel designs by decreasing variation among clusters and within clusters. Reducing variation does the equivalent of increasing the effect size. Covariates that decrease variation among clusters are particularly useful in multilevel designs because they decrease the effect of clustering by decreasing the ICC.

How Much Statistical Power Is Desirable?

Because statistical power is the probability of making the correct decision when a treatment effect actually exists, high statistical power is desirable. However, in any given design, higher power is achieved only with larger sample sizes. Obtaining larger sample sizes typically requires the commitment of more resources (not only costs but also research staff and burden on schools). Therefore, the benefits associated with higher power must be weighed against the commitment of resources required to achieve these benefits. Technical means alone cannot resolve this cost/benefit judgment. Normative statistical practice seems to be that power of 0.8 or above is considered acceptable, but there is no reason to think that this figure is always appropriate.

Chapter 5: Computing Statistical Power in Complex Designs

Specialized software is available for computing statistical power in group-randomized designs. One prominent example is Optimal Design, created by Stephen Raudenbush and his colleagues at the University of Michigan (Raudenbush et al. 2006). Although the Optimal Design software is useful, it does not cover all designs that are of interest to education researchers. For example, it does not cover designs using covariates at the individual level. Moreover, the use of software such as Optimal Design to compute power often fails to build intuition about how changes in parameters translate into changes in power. Finally, using Optimal Design properly requires a sophisticated understanding of the meaning of the input parameters. We hope that this paper will help education researchers build the intuition necessary to understand the conceptual meaning of parameters used in power analysis and how changes in design parameters translate into changes in power, whether they use Optimal Design or the methods described here.

Although specialized software is available, it is not necessary to obtain software in order to compute power for multilevel designs. Statistical power in cluster-randomized designs depends on sample sizes, intraclass correlation coefficients (ICCs), and covariate effects only through a design effect (similar in spirit—but not identical to—the variance inflation factor encountered in connection with two-stage clustered sampling). Each design has its own design effect. Except for the influence of the design effect, computation of statistical power in complex designs is much like computing power in simple designs that do not involve clustering.

Many tabulations (e.g., Cohen 1977) and computer programs (e.g., Borenstein, Rothstein, and Cohen 2001) are available for computing statistical power from designs involving simple random samples. The tables for computing power from the independent groups *t* test are perhaps the most widely available. Following Cohen's framework, such tables typically provide power values based on sample sizes n for each treatment group (assumed to be equal) and effect size δ (sometimes called Cohen's d):

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \qquad (2)$$

where μ_1 and μ_2 are the population means in the treatment and control groups, respectively, and σ is the within-group population standard deviation of the outcome. Table 1 is a slight variation on tables of this type. In Table 1, the statistical power of a two-sided test of the null hypothesis of no treatment effect at the $\alpha = 0.05$ level of significance is tabulated as a function of the total sample size N and δ. The only difference between Table 1 and the usual power tables, such as Table 2.3.2 on page 30 in Cohen (1977), is that N refers to the total sample size, rather than the sample size within each of two treatment groups. Thus, the row in Table 1 corresponding to any even value of N is equivalent to the row in Table 2.3.2 in Cohen with $n = N/2$. This slight modification is made for ease of use with certain of the operational sample sizes and operational effect sizes described below.

Computing Operational Sample and Effect Sizes

Tables 1 and 2 and tables like Cohen's (or the corresponding software) can be used to compute the power of the test used in the case of complex designs involving clustered sampling by appropriately adjusting the sample size and effect size used in the table. To use these tables to compute power for clustered designs, one must include in the table a total sample size N^T based on the number of clusters (here called the operational sample size) and a synthetic effect size Δ^T (here called the operational effect size) that will yield the appropriate power. Here the superscript T indicates that these quantities are used in the power tables. The design effect is different for each design, but the operational effect size will always be the product of the effect size (Cohen's d) and the design effect for that design:

$$\textit{operational effect size} = (\textit{effect size}) \times (\textit{design effect}) \qquad (3)$$

or

$$\Delta^T = \delta \times (\textit{design effect}).$$

Note that the term "design effect," as used here and in the rest of this paper, is not the same as the variance inflation factor encountered in connection with multistage cluster samples. This definition also differs from other definitions of design effects. The design effect in this context reflects the gain in precision obtained by having sampled more than one individual per cluster. Specifically, it is the square root of the ratio of the precision of the treatment effect estimate with one individual per cluster to the precision of the treatment effect estimate with n individuals per cluster. We note that the increase in precision that results from adding individuals per cluster will depend on the relevant ICCs or (in randomized-block designs) the extent to which treatment effects vary across clusters. A summary of design effects for hierarchical designs is given in appendix A and for randomized-block designs in appendix B.

Statistical power for multilevel designs can be computed without the use of tables, using functions in widely available statistical software. Using these functions to compute statistical power also involves using degrees of freedom based on the number of clusters and a simple function of the operational effect size (called the noncentrality parameter in statistics). The noncentrality parameter is different for each design, but it will always be the product of the operational effect size and a quantity that is a function of the sample size:

$$\textit{noncentrality parameter} = (\textit{operational effect size}) \times (\textit{sample size})$$

or

$$\lambda = \Delta^T \times (\textit{sample size}),$$

where (*sample size*) here is understood to mean some function of sample size specific to the design. The use of software functions to compute power makes it possible to avoid interpolation in tables and to automate the computation of a large number of power values (e.g., to examine many different design possibilities).

The sections that follow show how to use information about effect size, sample sizes at each level, ICC, and covariates to compute the statistical power of designs based on assigning schools, classrooms, or individuals to treatments. Each section indicates what the operational sample size is and how to compute the operational effect size for that design from the sample sizes and ICC and covariate information. How each factor influences statistical power and, therefore, how those factors might be manipulated to obtain a research design with adequate statistical power are shown. References are provided that may help in choosing plausible values of ICCs and correlations between covariates and outcomes for use in power calculations.

In each case, it is assumed that the experiment is planned to have a balanced design with equal numbers of individuals within each cluster. It is also assumed that the design is planned to have adequate power to compare two treatments (e.g., a treatment intervention and a control condition) because this is standard procedure even in designs that involve more than one active treatment. Finally, the term "school" will be used interchangeably with the term "cluster," and the term "class" or "classroom" will be taken to mean "subclusters" (in three-level designs) because these are the most likely terms to be used in education research and because the nesting relationship is readily understandable (it is common knowledge in this field that classrooms are nested within schools). The designations above are purely a matter of convenience: Nothing in this paper requires clusters to be schools or subclusters to be classrooms.

Chapter 6: Computing Power in Two-Level Hierarchical Designs That Assign Treatments to Clusters

Consider a two-level hierarchical experiment that will use a total of $2m$ clusters (typically schools) and will assign m of these schools to one treatment condition and m of these schools to another treatment condition (such as a control condition). Suppose that the sample size within each cluster has the same value n, so that mn individuals are assigned to each treatment and the total sample size is $N = 2mn$.

Suppose that the intervention effect at the population level is $(\mu_1 - \mu_2)$ in the units in which the outcome is measured (e.g., test score scale points). To compute statistical power, it is necessary to know the standardized intervention effect defined in equation (2), also known as the effect size. Suppose that the intraclass correlation coefficient (ICC) is ρ. Note that in discussions of two-level hierarchical designs, the symbol ρ, with no subscript, is used to denote ICC without ambiguity because there is only one possible intraclass correlation. In the discussion of three-level hierarchical designs, a subscript "S" or "C" is added to ρ to indicate the level (school or classroom) of the ICC.

Computing Power for Two-Level Hierarchical Designs With No Covariates

If the actual number of clusters assigned to each treatment is m, then the power table (Table 1) is entered with operational total sample size $N^T = 2m$. The operational effect size is

$$\Delta^T = \delta \sqrt{\frac{n}{1+(n-1)\rho}}, \tag{4}$$

where δ is the effect size, ρ is the ICC, and n is the sample size in each cluster. Note that unless $\rho = 1$, the design effect

$$\sqrt{\frac{n}{1+(n-1)\rho}}$$

is always larger than one, so the operational effect size Δ^T is always *larger* than the actual effect size δ. However, the operational sample size $N^T = 2m$ is *smaller* than the actual sample size $2mn$, so the power in the design with clustering will not be larger than in the design without clustering.

Using the operational effect size makes it possible to compute statistical power and sample size requirements for analyses based on clustered samples, using these tables or computer programs designed for the two-group t test. For example, entering Table 1 on the row given by the operational sample size N^T, and finding the column corresponding to the operational effect size Δ^T, one can read the power value.

Using a Computer Program to Compute Statistical Power in Two-Level Hierarchical Designs With No Covariates

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral *t*-distribution (Johnson and Kotz 1971). To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (related to the operational effect size) called the noncentrality parameter—specifically,

$$\lambda = \sqrt{\frac{m}{2}}\Delta^T = \delta\sqrt{\frac{mn}{2\left[1+\left(n-1\right)\rho\right]}}\,. \qquad (5)$$

If $H(x, v, \lambda)$ is the cumulative distribution function of the noncentral *t*-distribution with *v* degrees of freedom and noncentrality parameter λ, then the power of the one-tailed test for treatment effects at level α is

$$p_1 = 1 - H[c(\alpha, 2m-2), (2m-2), \lambda], \qquad (6)$$

where $c(\alpha, v)$ is the level α one-tailed critical value of the *t*-distribution with *v* degrees of freedom [e.g., $c(0.05,10) = 1.81$]. The power of the two-tailed test at level α is

$$p_2 = 1 - H[c(\alpha/2, 2m-2), (2m-2), \lambda] + H[-c(\alpha/2, 2m-2), (2m-2), \lambda]\,. \qquad (7)$$

The Impact of Design Parameters on Power for Two-Level Hierarchical Designs With No Covariates

This formulation helps make clear the effects of within-cluster sample size *n*, number of clusters per treatment *m*, ICC ρ, and effect size δ on statistical power. As Table 1 makes clear, for any fixed operational effect size, power increases rapidly as *m* increases, tending to 1.00 as *m* becomes large. Similarly, for any fixed operational sample size, power increases rapidly as δ increases, tending to 1.00 as δ becomes large. These are basic facts that also apply to a power analysis of simple designs.

The within-cluster sample size *n* and the ICC ρ affect power by altering the design effect. The impact of ρ on the design effect is the easiest to see. If $\rho = 0$, the design effect is $\sqrt{n}$, which is the maximum value of the design effect and, therefore, corresponds to the maximum operational effect size and the maximum power that can be attained in this design (where the values of *n* and δ are considered fixed). However, as ρ increases, the design effect, and therefore power, *decrease*. For example, when $n = 20$, the design effect is 4.47 when $\rho = 0.0$, but only 2.63 when $\rho = 0.1$; 2.04 when $\rho = 0.2$; 1.73 when $\rho = 0.3$; and, of course, the design effect is 1.0 when $\rho = 1.0$.

To see the impact of the within-cluster sample size *n* on the design effect, it is useful to rewrite the design effect as

$$\sqrt{\frac{n}{1+(n-1)\rho}} = \sqrt{\frac{1}{\frac{1}{n}+(1-\frac{1}{n})\rho}}.$$

This formulation makes clear that as n increases, the denominator of the design effect becomes smaller, so the design effect (and therefore the operational effect size) becomes larger, but only to a point. No matter how large n becomes, the design effect can never become larger than $\sqrt{1/\rho}$. Moreover, the design effect approaches this limiting value rather quickly. For example, if $\rho = 0.20$, the largest the design effect can be is $\sqrt{1/0.20} = 2.24$, but when $n = 10$, the design effect is already 1.89, and doubling n to $n = 20$ increases the design effect to only 2.04. Any further increases in n can have only very modest impacts on the design effect and therefore on power, demonstrating that, beyond a point (which occurs when n is rather modest), obtaining a larger sample size by increasing n has little effect on power.

Example. Consider a design for a study to evaluate the effects of a second-grade supplemental reading intervention in which whole schools have been assigned to either receive the treatment (intervention) or not. The effectiveness of the intervention will be measured by a post-treatment standardized reading test. In this intervention, recruitment comes from a broad range of schools, and there is some evidence (e.g., Hedges and Hedberg 2007) that a school-level ICC of about $\rho = 0.20$ is plausible. Note that this assumption is intended for illustration only and is not intended to suggest a value that will always be appropriate for other research studies. The study expects that at least 10 students will participate from each school, and previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$. The initial plan was for a study that would assign $m = 30$ schools to each condition. To determine the statistical power of this design, the operational effect size is first computed, using equation (4), as

$$\Delta^T = (0.35)\sqrt{\frac{10}{1+(10-1)0.20}} = (0.35)(1.89) = 0.661.$$

Entering Table 1 on the row corresponding to $N^T = 2m = 60$ shows that the power for $\Delta^T = 0.66$ will be between 0.63 (the power for $\Delta^T = 0.60$) and 0.76 (the power for $\Delta^T = 0.70$). Interpolating between these two values (0.66 is six-tenths of the way from 0.6 to 0.7, so six-tenths of the way between 0.63 and 0.76) obtains a power level of 0.71.

Because a power level of 0.71 is lower than desired, one might consider altering the design to increase power. For example, more students could be recruited from each school. Doing so would improve power only slightly. For example, if the number of students per school were increased by 50% to $n = 15$ (but the number of schools is kept constant at $m = 30$ per treatment group), the operational effect size would increase to only $\Delta^T = 0.695$, improving the power to only 0.75. If the number of students per school were doubled to $n = 20$ (but the number of schools were held constant at $m = 30$ per treatment group), the operational effect size would increase to only $\Delta^T = 0.714$, improving the power to only 0.77.

Increasing the number of schools has a much more dramatic effect on power. Increasing the number of schools by 50% to $m = 45$ per treatment group would have no effect on the

operational effect size, but power values from the row of the table for $N^T = 2m = 90$ for $\Delta^T = 0.60$ and 0.70 show that the power is between 0.80 and 0.91. Interpolating six-tenths of the way from 0.80 to 0.91 yields a power for $\Delta^T = 0.66$ of 0.87. An alternative way to increase power would be the use of covariates, which is discussed in the next section.

Two-Level Hierarchical Designs With Covariates

Now suppose that the analysis has q_S ($0 \le q_S < M - 2$) cluster-level covariates and q_W ($0 \le q_W < N - q_S - 2$) individual-level covariates.[1] For example, a design with $q_W = 1$ and $q_S = 1$ might arise if a pretest score (centered on the mean cluster score on the pretest) were used as an individual-level covariate and cluster means on the covariate were used as a group-level covariate. Note that individual-level covariates must be centered on cluster means for the power analyses described below to be exactly correct.

When covariates are used in the design both the operational effect size and the operational sample size need to be slightly modified. Table 1 is now entered with operational sample size $N_A^T = 2m - q_s$. This decrease in the operational sample size relative to a design without covariates reflects the degrees of freedom lost due to the modeling of between-cluster covariates.

The operational effect size is increased to an extent that depends on how much the covariates explain between- and within-cluster variance and how many cluster-level covariates are used. R_W^2 is the amount of within-cluster variance the covariates explain and R_S^2 is the amount of between-cluster variance the covariates explain. Thus, R_W^2 and R_S^2 can be thought of as proportions of variance (squared multiple correlations between the set of covariates and the response) accounted for in the usual way. The covariate-adjusted operational effect size is

$$\Delta_A^T = \delta \sqrt{\frac{2m}{2m - q_s}} \sqrt{\frac{n}{1 + (n-1)\rho - \left[R_W^2 + \left(nR_S^2 - R_W^2 \right) \rho \right]}} . \qquad (8)$$

Note that the covariate-adjusted design effect implied by equation (8) consists of two distinct parts. The first part, $\sqrt{2m/(2m - q_s)}$, is a correction term that depends on the sample size, taking into consideration the number of clusters ($2m$) and the number of cluster-level covariates q_s. This term is necessary because the number of degrees of freedom used in the t test depends on the number of cluster-level covariates modeled, but the noncentrality parameter λ does not depend on the number of covariates modeled. Note that the value of this factor is usually quite close to one. For instance, an experiment with $m = 20$ clusters assigned to each treatment group, using $q_s = 1$ cluster-level covariate, produces $\sqrt{2m/(2m - q_s)} = 1.013$, which differs from one by only about 1%.

The second part of the design effect,

[1] Note that the possibility of having 0 (no) covariates at a given level has been included in the previous section, and that adding covariates necessitates modifications of the operational sample size and the operational effect sizes.

$$\sqrt{\frac{n}{1+(n-1)\rho-\left[R_W^2+\left(nR_S^2-R_W^2\right)\rho\right]}},$$

is of a similar form to the unadjusted design effect. Because $\sqrt{2m/(2m-q_s)}$ is so often virtually one, the design effect will typically be quite close to the last factor in equation (8).

To compute power, enter Table 1 on the row given by the operational sample size $N_A{}^T = 2m - q_s$ and find the column corresponding to the operational effect size $\Delta_A{}^T$. The power value can then be read from the table.

To use the noncentral *t*-distribution function to compute the statistical power of a test, one must provide the program with a value of the covariate-adjusted noncentrality parameter,

$$\lambda_A = \sqrt{\frac{2m-q_S}{4}}\Delta_A^T = \delta\sqrt{\frac{mn/2}{1+(n-1)\rho-\left[R_W^2+\left(nR_S^2-R_W^2\right)\rho\right]}}. \qquad (9)$$

Power is computed using equation (6) for a one-tailed test or (7) for a two-tailed test, as above, except that the covariate-adjusted noncentrality parameter (9) is used with $2m - 2 - q_s$ degrees of freedom rather than $2m - 2$ degrees of freedom.

The Impact of Design Parameters on Power for Two-Level Hierarchical Designs With Covariates

As in the case without covariates, it is clear that—for any fixed operational effect size—power increases rapidly as *m* increases, tending to 1.00 as *m* becomes large. Similarly, for any fixed operational sample size, power increases rapidly as δ increases, tending to 1.00 as δ becomes large.

Moreover, the impact of the within-cluster sample size *n,* the ICC ρ, and the covariate-outcome (multiple) correlations (R_W within and R_S among clusters), occur entirely through the (second term in the) design effect. The effects of ρ and *n* on the design effect are similar to those in the design with no covariates. As ρ increases, the denominator of the design effect increases, the design effect decreases, and therefore power *decreases*. As *n* increases, the design effect (and therefore the operational effect size) becomes larger, but *n* has only a very modest impact on the design effect (and therefore on power) beyond a certain point. Thus, as in the design without covariates, beyond a point (which occurs when *n* is rather modest), obtaining a larger sample size by increasing *n* has little impact on power. However, as *n* becomes large in the design with covariates, the maximum design effect is now

$$\sqrt{\frac{2m}{2m-q_S}}\sqrt{1/(1-R_S^2)\rho}$$

instead of $\sqrt{1/\rho}$, as it was in the design without covariates.

Note that the denominator of the design effect has two terms. The first term,

$$1 + (n - 1)\rho,$$

is the denominator of the design effect in the design without covariates. The second term,

$$-\left[R_W^2 + \left(nR_S^2 - R_W^2\right)\rho\right],$$

has the same form as $[1 + (n - 1)\rho]$, except that R_W^2 replaces 1 and nR_S^2 replaces n. Because R_S^2 and R_W^2 are typically larger than zero (and cannot be smaller than zero), this second term is negative and increases the design effect. If $R_S^2 = R_W^2 = 0$, the covariates have no effect and the design effect becomes the same as in the design with no covariates.

The presence of q_S in the denominator of equation (8) at first glance suggests that power can be increased simply by using more cluster-level covariates. This advantage is illusory because although a larger value of q_S will make Δ_A^T larger, it will also make the operational sample size, N_A^T, smaller. In fact, unless the addition of more cluster-level covariates increases the value of R_S^2, this addition can only harm power, not help it.

Example. Return to the design considered earlier for a study to evaluate the effects of a second-grade supplemental reading intervention where a school-level ICC of about $\rho = 0.20$ is plausible, $n = 10$ students would participate from each school, and previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$. Suppose that pretreatment reading test scores are available and that both the individual pretest scores and the school means for this pretest will be used as covariates in the analysis. Thus, $q_s = 1$ and $q_w = 1$. There is some evidence that values of $R_W^2 = 0.5$ and $R_S^2 = 0.8$ are plausible (see Table 3 in Hedges and Hedberg 2007), although we caution, as we did in the previous example, that these values are mainly intended for illustration and should not necessarily be interpreted as reference values to use when planning other research studies. The initial plan was for a study that would assign $m = 20$ schools to each condition. To determine the statistical power of this design, the operational effect size is computed using equation (8) as

$$\Delta_A^T = (0.35)\sqrt{\frac{40}{40-1}}\sqrt{\frac{10}{1+(10-1)(0.2)-\left[0.5+(10\times 0.8-0.5)0.2\right]}} = (0.35)(1.013)(3.536) = 1.253$$

Entering Table 1 on the row corresponding to $N_A^T = 2m - q_s = 39$ yields a power for $\Delta^T = 1.25$ that will be between 0.95 (the power for $\Delta^T = 1.20$) and 0.98 (the power for $\Delta^T = 1.30$). Interpolating between these two values (1.25 is halfway from 1.2 to 1.3) requires going half of the way between 0.95 and 0.98, which yields a power level of 0.965, or 0.96 (to round lower and be slightly conservative).

This high statistical power might be seen as providing a margin of safety in case any assumptions are somewhat optimistic. Alternatively, because a power level of 0.96 may be higher than necessary, one might consider altering the design to decrease costs while maintaining acceptable statistical power. For example, one might consider decreasing the number of schools to $m = 15$ per treatment group. That would have very little effect on the operational effect size (it would now be 1.259), but reading power values from the row of the table for $N_A^T = 2m - q_s = 29$ for $\Delta^T = 1.20$ and 1.30 shows that the power is between 0.88 and 0.92. Interpolating six-tenths of the way from 0.89 and 0.93 gives a power for $\Delta^T = 1.24$ of 0.90. Comparing this value with that derived in the last section for the same study without covariates, one sees that a design with $m = 15$ schools using a pretest as a covariate has greater power than a design with $m = 45$ schools (three times as many schools) using no covariates (assuming the R^2 values used are accurate).

Chapter 7: Computing Statistical Power in Three-Level Hierarchical Designs

Consider a three-level hierarchical experiment that will use a total of *2m* clusters (typically schools) and will assign *m* of these schools to a treatment condition and *m* of these schools to a control condition. Suppose that each school has a total of *p* subclusters (usually classrooms) and that the sample size within each classroom has the same value *n*, so that the total sample size is $N = 2mpn$.

Suppose that the intervention effect at the population level is $(\mu_1 - \mu_2)$ in the units in which the outcome is measured (e.g., test score scale points). In three-level designs, statistical power also depends on the intervention effect via the effect size or standardized intervention effect (sometimes called Cohen's *d*):

$$\delta = \frac{\mu_1 - \mu_2}{\sigma_T},$$

where σ_T is the total population standard deviation of the outcome.

In three-level models, two indices are necessary to characterize the relationship between the component variances that make up the total variance, and they are generalizations of the intraclass correlation coefficient (ICC). Let $\sigma_T^2 = \sigma_S^2 + \sigma_C^2 + \sigma_W^2$ be the total variance, where σ_S^2 is the between-cluster (e.g., between-school) variance, σ_C^2 is the between-subcluster but within-cluster (e.g., between-classroom within-school) variance, and σ_W^2 is the within-subcluster (e.g., within-classroom) variance. Define the school-level ICC ρ_S by

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2} = \frac{\sigma_S^2}{\sigma_T^2}. \qquad (10)$$

Similarly, define the classroom-level ICC ρ_C by

$$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2} = \frac{\sigma_C^2}{\sigma_T^2}. \qquad (11)$$

Together, the ICCs ρ_S and ρ_C define the clustering structure in the three-level experiment.

Three-Level Hierarchical Designs With No Covariates

Using the same ideas as for two-level hierarchical designs, one can compute power for three-level hierarchical designs by using the appropriate operational sample size and operational effect size.

Because the actual number of clusters assigned to each of the two treatments is m, we enter the power table with operational total sample size $N^T = 2m$. The operational effect size is

$$\Delta^T = \delta \sqrt{\frac{pn}{1+(pn-1)\rho_S + (n-1)\rho_C}}, \tag{12}$$

where δ is the effect size, ρ_S is the cluster-level (school) ICC, ρ_C is the subcluster-level (classroom) ICC, p is the number of subclusters (classrooms) per cluster (school), and n is the number of individuals in each subcluster (classroom). Note that if

$$\frac{pn-1}{n-1} > \frac{\rho_c}{1-\rho_s},$$

the design effect is greater than one. Because it will generally be the case that $\rho_c < 0.5$ and $\rho_s < 0.5$, the design effect will usually be greater than one, so the operational effect size Δ^T is usually larger than the actual effect size δ. However, the operational total sample size N^T is smaller than the actual total sample size $2mpn$, so the power in the design with clustering will be smaller than in the design without clustering.

Using the operational effect size makes it possible to compute statistical power and sample size requirements for analyses based on clustered samples using tables and computer programs designed for the independent groups t test. For example, one can read the power value by entering Table 1 on the row given by the operational sample size N^T and finding the column corresponding to the operational effect size Δ^T.

Computing Statistical Power for Noncentral t-Distributions Using Computer Programs

An alternative method for computing statistical power in three-level hierarchical designs is to use a computer program that has a built-in function that computes the noncentral t-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of the noncentrality parameter—specifically,

$$\lambda = \sqrt{\frac{m}{2}}\Delta^T = \delta \sqrt{\frac{mpn}{2\left[1+(pn-1)\rho_S + (n-1)\rho_C\right]}}. \tag{13}$$

Power is computed using equation (6) for a one-tailed test or (7) for a two-tailed test, as above. Note that the degrees of freedom used is the same as in the two-level hierarchical design; that is, the degrees of freedom do not depend on the number of subclusters sampled (Konstantopoulos 2008b).

The Impact of Design Parameters on Power for Three-Level Hierarchical Designs With No Covariates

This formulation helps make clear the effects of the number of clusters *m*, the number of subclusters *p*, within-subcluster sample size *n*, ICCs ρ_S and ρ_C, and effect size δ on statistical power. As in the case of two-level designs, for any fixed operational effect size, power increases rapidly as *m* increases, tending to 1.00 as *m* becomes large. Similarly, for any fixed operational sample size, power increases rapidly as δ increases, tending to 1.00 as δ becomes large.

The impact of the number of subclusters per cluster *p*, the within-subcluster sample size *n,* and the ICCs ρ_S and ρ_C occurs entirely through the design effect. The impact of ρ_S and ρ_C on the design effect is the easiest to see. If $\rho_S = \rho_C = 0$, the design effect is $\sqrt{pn}$, which is the maximum value of the design effect and therefore corresponds to the maximum operational effect size and the maximum power that can be attained in this design. However, as either ρ_S or ρ_C increases, the design effect—and therefore power—*decreases*. Furthermore, an increase in ρ_S will have a more deleterious effect on power than will a similarly sized increase in ρ_c. For example, when $n = 20$ and $p = 3$, the design effect is 7.75 when $\rho_s = \rho_c = 0$; 4.55 when $\rho_s = 0$ and $\rho_c = 0.1$; 2.99 when $\rho_s = 0$ and $\rho_c = 0.3$; and 1.73 when $\rho_s = 0$ and $\rho_c = 1$. Similarly, when $\rho_c = 0$ and $\rho_s = 0.1$, the design effect decreases from 7.75 to 2.95; when $\rho_c = 0$ and $\rho_s = 0.3$, the design effect is 1.79; and when $\rho_c = 0$ and $\rho_s = 1$, the design effect is, of course, one.

To see the impact of the within-subcluster sample size *n* on the design effect, it is useful to rewrite the design effect as

$$\sqrt{\frac{pn}{1+(pn-1)\rho_S+(n-1)\rho_C}} = \sqrt{\frac{1}{\frac{1}{pn}+(1-\frac{1}{pn})\rho_S+(\frac{1}{p}-\frac{1}{pn})\rho_C}}\,.$$

This formulation makes clear that as *n* increases, the denominator of the design effect (as expressed on the right, above) becomes smaller, so the design effect (and therefore the operational effect size) becomes larger, but only to a point. No matter how large *n* becomes, the design effect can never become larger than

$$\sqrt{\frac{1}{\rho_S+\frac{1}{p}\rho_C}}\,.$$

Moreover, the design effect approaches this limiting value rather quickly. For example, if $\rho_S = 0.20$, $\rho_C = 0.10$, and $p = 2$, the largest the design effect can be is $\sqrt{1/[0.20+(0.10/2)]} = 2.00$, but when $n = 10$, the design effect is already 1.87, and doubling *n* to $n = 20$ increases the design effect to only 1.93. Any further increases in *n* can have only a very modest impact on the design effect and therefore on power. This explication demonstrates that beyond a point (which occurs when *n* is rather modest), obtaining a larger sample size by increasing *n* has little impact on power.

Example. Return to the design considered earlier for a study to evaluate the effects of a second-grade supplemental reading intervention. Remember that the assumptions made here are intended

for illustration only and are not intended to suggest values that will always be appropriate for other research studies. Continue to assume that a school-level ICC of about $\rho_S = 0.20$ is plausible and assume that a classroom-level ICC of roughly $\rho_c = 0.13$ is plausible. Assume, in addition, that $n = 10$ students from each of $p = 2$ classrooms would participate from each school. Previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$. The initial plan was for a study that would assign $m = 30$ schools to each condition. To determine the statistical power of this design, first compute the operational effect size using equation (12):

$$\Delta^T = (0.35)\sqrt{\frac{2(10)}{1+[2(10)-1](0.20)+(10-1)(0.13)}} = (0.35)(1.830) = 0.641.$$

Entering Table 1 on the row corresponding to $N^T = 2m = 60$, one sees that the power for $\Delta^T = 0.64$ will be between 0.63 (the power for $\Delta^T = 0.60$) and 0.76 (the power for $\Delta^T = 0.70$). Interpolating between these two values (0.64 is four-tenths of the way from 0.6 to 0.7, going four-tenths of the way between 0.63 and 0.76) yields a power level of 0.68.

Because a power level of 0.68 is lower than desired, one might consider altering the design to increase power. Similar to the case of the two-level hierarchical design, neither increasing the number of students, n, per classroom, nor increasing the number of classrooms, p, per school (even if feasible) would have a substantial effect on statistical power.

Increasing the number of schools, however, has a much more dramatic effect on power. Increasing the number of schools by 50% to $m = 45$ per treatment group would have no effect on the operational effect size, but power values from the row of the table for $N^T = 2m = 90$ for $\Delta^T = 0.60$ and 0.70 show that the power is between 0.80 and 0.91. Interpolating four-tenths of the way from 0.80 to 0.91, one gets a power of approximately 0.84 for $\Delta^T = 0.64$. Using covariates also can increase power; this method is discussed in the next section.

Three-Level Hierarchical Designs With Covariates

Now suppose that there are q_S ($0 \le q_S < 2m - 2$) cluster-level covariates, q_C ($0 \le q_C < 2mp - q_S - 2$) subcluster-level covariates, and q_W ($0 \le q_W < N - q_S - q_C - 2$) individual-level covariates in the analysis.[2] For example, a design with $q_W = 1$, $q_C = 1$, and $q_S = 1$ might arise if a pretest were used (centered on subcluster means) as an individual-level covariate; subcluster means (centered on cluster means) on the pretest were used as a subcluster-level covariate; and cluster means on the pretest were used as a cluster-level covariate. The centering of covariates on higher-level means is again crucial for the power computations described below to be exact.

When covariates are used in the design, both the operational effect size and the operational sample size are slightly modified. As with the two-level design, operational sample size $N_A^T = 2m - q_s$ is entered into table 1.This decrease in the operational sample size relative to a

[2] Note that the possibility of having 0 (no) covariates at a given level has been included in the previous section, and that adding covariates necessitates modifications of the operational sample size and the operational effect sizes.

design without covariates reflects the degrees of freedom lost due to the modeling of between-cluster covariates.

The operational effect size increases to an extent that depends on how much the covariates explain between-cluster, between-subcluster but within-cluster, and within-subcluster variance. R_W^2 is the amount of within-subcluster variance the covariates explain, R_S^2 is the amount of between-cluster (between-school) variance the covariates explain, and R_C^2 is the amount of between-subcluster but within-cluster (between-classroom but within-school) variance the covariates explain. One can think of R_W^2, R_S^2, and R_C^2 as proportions of variance accounted for (squared multiple correlations) in the usual way. The covariate-adjusted operational effect size is

$$\Delta_A^T = \delta \sqrt{\frac{2m}{2m - q_s}} \sqrt{\frac{pn}{1 + (pn-1)\rho_S + (n-1)\rho_C - \left[R_W^2 + (pnR_S^2 - R_W^2)\rho_S + (nR_C^2 - R_W^2)\rho_C \right]}} \quad (14)$$

In complete analogy with the two-level hierarchical design, the covariate-adjusted design effect implied by equation (14) consists of two distinct parts: a correction term depending on the operational sample size and a second term that contains the information about the effects of clustering and adjustment for covariates on the operational effect size. Because the first factor is again generally quite close to one, the design effect will be quite close to the last factor in equation (14).

Because the term in square brackets in the denominator of the (second term in the) design effect is never less than zero and is generally positive, the covariate-adjusted operational effect size Δ_A^T is generally larger than the unadjusted operational effect size Δ^T. In computing Δ_A^T, it may be more convenient to break the denominator of the (second term in the) design effect into two parts. One part

$$A = 1 + (pn - 1)\rho_S + (n - 1)\rho_C$$

reflects the impact of clustering, and the second part

$$B = R_W^2 + (pnR_S^2 - R_W^2)\rho_S + (nR_C^2 - R_W^2)\rho_C$$

reflects the adjustment for the effects of covariates, so that

$$\Delta_A^T = \delta \sqrt{\frac{2m}{2m - q_s}} \sqrt{\frac{pn}{A - B}}. \quad (15)$$

Entering Table 1 on the row given by the operational sample size N_A^T and finding the column corresponding to the operational effect size Δ_A^T, one can read the power value.

To use the noncentral *t*-distribution function to compute the statistical power of a test, one must provide the program with a value of the covariate-adjusted noncentrality parameter,

$$\lambda_A = \sqrt{\frac{2m - q_S}{4}}\Delta_A^T = \delta\sqrt{\frac{mpn/2}{1+(pn-1)\rho_S+(n-1)\rho_C-\left[R_W^2+(pnR_S^2-R_W^2)\rho_S+(nR_C^2-R_W^2)\rho_C\right]}}$$

Power is computed using equation (6) for a one-tailed test or (7) for a two-tailed test, as above, except that the covariate-adjusted noncentrality parameter (16) is used with $2m - q_s - 2$ degrees of freedom.

The Impact of Design Parameters on Power for Three-Level Hierarchical Designs With Covariates

As in the case without covariates, it is clear that for any fixed operational effect size, power increases rapidly as m increases, tending to 1.00 as m becomes large. Similarly, for any fixed operational sample size, power increases rapidly as δ increases, tending to 1.00 as δ becomes large.

Moreover, the impact on power of the number of subclusters p within each cluster; the within-subcluster sample size n; the ICCs, ρ_S, and ρ_C; and the covariate-outcome (multiple) correlations R_W, R_S, and R_C, occurs entirely through the design effect. The impact of ρ_S, ρ_C, p, and n on the design effect is essentially the same as in the design with no covariates. As ρ_S and ρ_C increase, the denominator of the design effect increases, the design effect decreases, and therefore power *decreases*. As p and n increase, the design effect (and therefore the operational effect size) becomes larger, but beyond a certain point, p and n have only a very modest impact on the design effect and therefore on power. Thus, as in the design without covariates, obtaining a larger sample size by increasing p and n beyond this point (which occurs when p and n are rather modest) has little impact on power.

The impact of covariates on power is easier to understand. Note that the denominator of the (second term in the) design effect has two terms. The first term,

$$1 + (pn - 1)\rho_S + (n - 1)\rho_C,$$

is the denominator of the design effect in the design without covariates. The second term of the denominator of the design effect,

$$- [R_W^2 + (pnR_S^2 - R_W^2)\rho_S + (nR_C^2 - R_W^2)\rho_C],$$

has the same form as $1 + (pn - 1)\rho_S + (n - 1)\rho_C$, except that R_W^2 replaces 1, pnR_S^2 replaces pn, and nR_C^2 replaces n. Because R_W^2, R_S^2, and R_C^2 are typically larger than zero (and are never smaller than zero), this second term is negative and increases the design effect. If $R_S^2 = R_C^2 = R_W^2 = 0$, the covariates have no effect, and the design effect becomes the same as in the design with no covariates.

Example. Return to the design for a study to evaluate the effects of a second-grade supplemental reading intervention where a school-level ICC of about $\rho_s = 0.20$ is plausible and a classroom-level ICC of about $\rho_c = 0.13$ is plausible, $n = 10$ students would participate from each of $p = 2$

classes in each school, and previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$. Suppose that pretreatment reading test scores are available and that classroom-centered individual pretest scores, school-centered classroom mean pretest scores, and school mean pretest scores will be used as covariates. Thus, $q_W = q_C = q_S = 1$. There is some evidence that values of $R_W^2 = 0.5$, $R_C^2 = 0.6$, and $R_S^2 = 0.8$ are plausible (Hedges and Hedberg 2007). These values are again mainly intended for illustration and should not necessarily be interpreted as reference values to use when planning other research studies.
The initial plan was for a study that would assign $m = 30$ schools to each condition. To determine the statistical power of this design, one first computes the operational effect size, using equation (15), with

$$A = 1 + [\,(2)(10) - 1](0.2) + (10 - 1)(0.13) = 5.970$$

and

$$B = 0.5 + [(2)(10)(0.8) - 0.5](0.2) + [(10)(0.6) - 0.5](0.13) = 4.315,$$

so that

$$\Delta_A^T = 0.35\sqrt{\frac{60}{59}}\sqrt{\frac{2(10)}{5.970 - 4.315}} = (0.35)(1.01)(3.476) = 1.227\,.$$

Entering Table 1 on the row corresponding to $N^T = 2m = 60$, one sees that the power for $\Delta^T = 1.2$ is listed as 1.00, so the power for $\Delta^T = 1.23$ is at least 0.995.

One might regard this high statistical power as providing a margin of safety in case any assumptions are somewhat optimistic. Alternatively, because power of 0.995 may be higher than necessary, one might consider altering the design to decrease costs while maintaining acceptable statistical power. For example, one might consider decreasing the number of schools to $m = 15$ per treatment group. That decrease would have very little effect on the operational effect size (it would change to 1.238), but reading power values from the row of Table 1 for $N^T = 2m\text{-}q = 30$ for $\Delta^T = 1.2$, one sees that the power is at least 0.89.

Comparing this value with that derived in the last section for the same study without covariates, we see that a design with $m = 15$ schools, using a pretest as a covariate, has higher power than a design with $m = 45$ schools (three times as many schools), using no covariates (again, assuming the R^2 values used are accurate).

Chapter 8: Computing Power in Randomized-Block Designs

Computing Power in Two-Level Randomized-Block Designs That Assign Treatments Within Clusters

Consider a two-level experiment that uses a total of m clusters (typically schools), with $2n$ individuals in each cluster. Unlike the hierarchical design that assigns whole clusters (e.g., schools) to treatments, the experiment assigns some individuals within each cluster to each of two treatments. That is, within each of the m schools, n individuals are assigned to each treatment, so that mn individuals are assigned to each treatment and the total sample size is $N = 2mn$.

Suppose that the intervention effect at the population level is $(\mu_1 - \mu_2)$ in the units in which the outcome is measured (e.g., test score scale points). Rather than this unstandardized effect, statistical power is computed on the basis of the effect size or standardized intervention effect (sometimes called Cohen's d):

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where σ is the total population standard deviation of the outcome within treatment groups. That is, $\sigma^2 = \sigma_S^2 + \sigma_W^2$.

In the randomized-block design, as in the hierarchical design, power depends on the cluster-level intraclass correlation coefficient (ICC), $\rho = \sigma_S^2 / (\sigma_S^2 + \sigma_W^2)$. However, in this design, power also depends on the degree to which treatment effects vary across clusters. It is convenient to characterize this treatment effect heterogeneity via the parameter ω, which represents the proportion of between-cluster variability that is attributable to heterogeneity of treatment effects. Thus, ω can be characterized as

$$\omega = \sigma_{T\times S}^2 / \sigma_S^2, \qquad (17)$$

where $\sigma_{T\times S}^2$ is the variance due to the treatment by cluster interaction and σ_S^2 is the total cluster level variance. When the usual statistical model is used for power analysis, under very mild additional assumptions $\sigma_{T\times S}^2$ is less than σ_S^2, so that ω is almost always less than one and can be as small as zero in cases where the treatment effect is very similar across clusters. For example, a class size experiment involved kindergarten through fourth-grade students in 79 elementary schools in 42 school districts in Tennessee. The experiment had a randomized-block design, assigning treatments to classrooms within schools. Nye, Hedges, and Konstantopoulos (2000) estimated the between-school variance of small-class effects and found that variation of treatment effects was reasonably small for most grades and subject matter (with an average ω value of about 0.3). However, it is possible for ω to be larger, particularly when treatment implementation may vary widely.

Two-Level Randomized-Block Designs With No Covariates

If the experiment has m clusters, enter the one-sample t test power table with operational sample size $N^T = m$. The operational effect size is

$$\Delta^T = \delta \sqrt{\frac{n}{2\left[1+\left(n\omega-1\right)\rho\right]}}, \tag{18}$$

where δ is the effect size, ρ is the ICC, ω is one half of the ratio of variance of the treatment effects across clusters to the total cluster level variance, and n is the sample size in each cluster. Note that the design effect

$$\sqrt{\frac{n}{2\left[1+\left(n\omega-1\right)\rho\right]}}$$

can be larger than one, so the operational effect size Δ^T can be larger than the actual effect size δ. However, the operational sample size $N^T = m$ is smaller than the actual within-treatment sample size mn, so the power in the design with clustering will usually not be larger than in the design without clustering.

Comparing the operational effect size in the randomized-block design with that in the two-level hierarchical designs having the same total sample size, and noting that ω is usually smaller than one, one can see that the operational effect size Δ^T is usually larger in the randomized-block design.

Using the operational effect size makes it possible to compute statistical power and sample size requirements for analyses based on clustered samples using tables and computer programs designed for the one-sample t test. For example, entering Table 2 on the row given by the operational sample size N^T, and finding the column corresponding to the operational effect size Δ^T, one can read the power value.

Using a Computer Program to Compute Statistical Power in Two-Level Randomized-Block Designs With No Covariates.

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral t-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (related to the operational effect size) called the noncentrality parameter, specifically

$$\lambda = \sqrt{m}\Delta^T = \delta \sqrt{\frac{mn}{2\left[1+\left(n\omega-1\right)\rho\right]}}. \tag{19}$$

If H(x, v, λ) is the cumulative distribution function of the noncentral t-distribution with v degrees of freedom and noncentrality parameter λ, then the power of the one-tailed test for treatment effects at level α is

$$p_1 = 1 - \mathrm{H}[c(\alpha, m-1), (m-1), \lambda], \tag{20}$$

where $c(\alpha, v)$ is the level α one-tailed critical value of the t-distribution with v degrees of freedom [e.g., $c(0.05,10) = 1.81$]. The power of the two-tailed test at level α is

$$p_2 = 1 - \mathrm{H}[c(\alpha/2, m-1), (m-1), \lambda] + \mathrm{H}[-c(\alpha/2, m-1), (m-1), \lambda]\ . \tag{21}$$

These results are given in a slightly different notation in Raudenbush and Liu (2000).

The Impact of Design Parameters on Power in Two-Level Randomized-Block Designs With No Covariates.

This formulation helps to clarify the effects of within-cluster sample size $2n$, number of clusters m, ICC ρ, the heterogeneity parameter ω, and effect size δ on statistical power. As Table 2 shows, for any fixed operational effect size, power increases rapidly as m increases, tending to 1.00 as m becomes large. Similarly, for any fixed operational sample size, power increases rapidly as δ increases and tends to 1.00 as δ becomes large. These are basic facts from the power analysis of simple designs.

The within-cluster sample size n, the ICC ρ, and the heterogeneity parameter ω impact power because they impact the design effect. The effect of the heterogeneity parameter can be profound. If treatment effects are perfectly consistent across clusters so that $\omega = 0$, the design effect is $\sqrt{n/[2(1-\rho)]}$, which is the maximum value of the design effect for fixed values of n and ρ and therefore corresponds to the maximum operational effect size and the maximum power that can be attained in this design. However, as the heterogeneity of treatment effects increases, the design effect (and therefore power) decreases.

To see the impact of the within-cluster sample size n on the design effect, it is useful to rewrite the design effect as

$$\sqrt{\frac{n/2}{1+(n\omega-1)\rho}} = \sqrt{\frac{1/2}{\frac{1}{n}+(\omega-\frac{1}{n})\rho}}\ .$$

This formulation makes clear that as n increases, the denominator of the design effect becomes smaller, so the design effect (and therefore the operational effect size) becomes larger, but only to a point. No matter how large n becomes, the design effect can never become larger than $\sqrt{1/[2\omega\rho]}$. Moreover, the design effect approaches this limiting value rather quickly. For example, if $\rho = 0.20$ and $\omega = 0.5$, the largest the design effect can be is $\sqrt{1/[2(0.20)0.5]} = 2.24$, but when $n = 10$, the design effect is already 1.67 and doubling n to $n = 20$ increases the design effect to only 1.89. Any further increases in n can have only a very modest impact on the design

effect and therefore on power, demonstrating that, beyond a point (which occurs when n is rather modest), obtaining a larger sample size by increasing n has little impact on power.

Example. Return to the problem of designing a study to evaluate the effects of a second-grade supplemental reading intervention. Suppose that it is reasonable to believe that this supplemental reading intervention might be administered to some students and not to others in the same school without fear of contamination. Thus, a two-level randomized-block design is used. It is still the intention to recruit from a broad range of schools so that a school-level ICC of about $\rho = 0.20$ is plausible (e.g., Hedges and Hedberg 2007). One expects that at least $n = 10$ students will participate in each treatment group from each school, and previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$ and that effects are likely to be fairly consistent across schools, so that a value of $\omega = 0.5$ is plausible (even fairly conservative). The initial plan was for a study that would involve $m = 30$ schools. Recall that these assumptions are intended for illustration only and are not intended to suggest values that will always be appropriate for other research studies.

To determine the statistical power of this design, one first computes the operational effect size, using equation (18):

$$\text{€}^T = (0.35)\sqrt{\frac{10}{2\left\{1+\left[(10)(0.5)-1\right]0.20\right\}}} = (0.35)(1.667) = 0.583\,.$$

Entering Table 2 on the row corresponding to $N^T = m = 30$, one sees that the power for $\Delta^T = 0.58$ will be between 0.75 (the power for $\Delta^T = 0.50$) and 0.89 (the power for $\Delta^T = 0.60$). Interpolating between these two values (0.58 is eight-tenths of the way from 0.5 to 0.6, so one needs to go eight-tenths of the way between 0.75 and 0.89), one obtains a power level of 0.86.

Note that the power of a two-level hierarchical design that assigned the same number of students $mn = (30)(10) = 300$ to each treatment group (but used twice as many schools because there were half as many individuals per school) was considerably lower (only 0.71). This example illustrates how much higher the power of a randomized-block design may be if the design can be used.

Note that this power calculation is somewhat sensitive to the value of the parameter ω that describes the heterogeneity of treatment effects across clusters. If ω had been twice as large (that is, $\omega = 1.0$), the operational effect size would have been only $\Delta^T = 0.47$, and the power would have been only approximately 0.69.

Two-Level Randomized-Block Designs With Covariates

Now suppose that there are q_S ($0 \le q_S < m - 1$) cluster-level covariates and q_W ($0 \le q_W < N - q_S - 1$) individual-level covariates in the analysis.[3] For example, a design with $q_W = 1$ and $q_S = 1$ might arise if a pretest were used as an individual-level covariate (centered on cluster means) and cluster means on the covariate were used as a group-level covariate.

When covariates are used in the design, both the operational effect size and the operational sample size are slightly modified. As was the case in hierarchical designs, the operational sample size is decreased relative to a design without covariates to reflect the degrees of freedom lost due to the modeling of between-cluster covariates. In the two-level randomized-block design, one enters Table 2 with operational sample size

$$N_A^T = m - q_s .$$

As was the case in hierarchical designs, the operational effect size is increased when covariates are used, but the nature of this modification is somewhat different. In the randomized-block design, cluster-level covariates increase power primarily if they explain part of the variance in treatment effects across clusters (that is, if they explain part of the cluster by treatment interaction variance). Contrary to the hierarchical design, cluster-level covariates that explain only variation among cluster means will have no effect on power in the randomized-block design. R_W^2 is the amount of within-cluster variance the covariates explain, and R_{TS}^2 is the amount of between-cluster variance *in treatment effects* the covariates explain. One can think of R_W^2 and R_{TS}^2 as proportions of variance accounted for (squared correlations) in the usual way. The covariate-adjusted operational effect size is

$$\Delta_A^T = \delta \sqrt{\frac{m}{m - q_S}} \sqrt{\frac{n/2}{1 + (n\omega - 1)\rho - \left[R_W^2 + \left(n\omega R_{TS}^2 - R_W^2 \right) \rho \right]}} . \qquad (22)$$

Note that the covariate-adjusted design effect implied by equation (22) consists of two distinct parts. The first part, $\sqrt{m/(m - q_s)}$, is a correction term that depends on the sample size (number of clusters m) and the number of cluster-level covariates q_s. This term is necessary because the degrees of freedom used in the t test depend on the number of cluster-level covariates modeled, but the noncentrality parameter λ does not depend on the number of covariates modeled. Note that the value of this factor is usually quite close to one. For instance, in an experiment with m = 40 total clusters and q_s = 1 cluster-level covariate used, $\sqrt{m/(m - q_s)} = 1.013$, so this factor differs from one by only about 1%.

The second part of the design effect,

$$\sqrt{\frac{n/2}{1 + (n\omega - 1)\rho - \left[R_W^2 + \left(n\omega R_{TS}^2 - R_W^2 \right) \rho \right]}} ,$$

[3] Note that the possibility of having 0 (no) covariates at a given level has been included in the previous section, and that adding covariates necessitates modifications of the operational sample size and the operational effect sizes.

is of a similar form to the unadjusted design effect Δ^T. Because $\sqrt{m/(m-q_s)}$ is so often virtually one, the design effect will typically be quite close to the last factor in equation (22).

The covariate-adjusted operational effect size is generally larger than the unadjusted design effect Δ^T because R_{TS}^2 and R_W^2 are generally positive; hence, the term in square brackets in the denominator of the second term of the design effect is usually negative.

Using the operational effect size makes it possible to compute statistical power and sample size requirements for analyses based on clustered samples using the tables and computer programs designed for the one sample *t* test. For example, entering Table 2 on the row given by the operational sample size N^T, and finding the column corresponding to the operational effect size Δ^T, one can read the power value.

Using a Computer Program to Compute Statistical Power in Two-Level Randomized-Block Designs With Covariates.

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral *t*-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (associated with the operational effect size) called the noncentrality parameter—specifically,

$$\lambda_A = \sqrt{m-q_S}\,\Delta_A^T = \delta\sqrt{\frac{mn/2}{1+(n\omega-1)\rho-\left[R_W^2+\left(n\omega R_{TS}^2-R_W^2\right)\rho\right]}}. \qquad (23)$$

Power is computed using equation (20) for a one-tailed test or (21) for a two-tailed test, as above, except that the covariate-adjusted noncentrality parameter (23) is used with $m - 1 - q_S$ degrees of freedom (Raudenbush and Liu 2000).

Example. Returning to the problem of designing a study to evaluate the effects of a second-grade supplemental reading intervention, suppose that it is reasonable that this intervention might be administered to some students and not to others in the same school, without fear of contamination. In this case, a two-level randomized-block design would be used. One would still intend to recruit from a broad range of schools so that a school-level ICC of about $\rho = 0.20$ is plausible (Hedges and Hedberg 2007). Again, suppose that at least $n = 10$ students would participate in each treatment group from each school. Previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$ and that effects are likely to be fairly consistent across schools, so that a value of $\omega = 0.5$ is plausible (even fairly conservative). We continue to assume that the values $R_W^2 = 0.5$ and $R_S^2 = 0.8$ are plausible. We also assume that about half of the total variance between schools that the covariates explain is due to the ability of the covariates to predict treatment effect variability, so that $R_{TS}^2 = 0.4$. We again clarify that these assumptions are intended for illustration only and are not intended to suggest values that will always be appropriate for other research studies. The initial plan was for a study that would

involve $m = 30$ schools. To determine the statistical power of this design, one first computes the operational effect size, using equation (22):

$$\Delta_A^T = (0.35)\sqrt{\frac{30}{29}}\sqrt{\frac{10/2}{1+\left[(10)(0.5)-1\right](0.2)-\{0.5+[(10)(0.5)(0.4)-0.5](0.2)\}}}\,.$$
$$= (0.35)(1.017)(2.236) = 0.80$$

Entering Table 2 on the row corresponding to $N^T = m - 1 = 29$, one sees that the power for $\Delta^T = 0.80$ will be indistinguishable from 0.99.

Computing Statistical Power in Three-Level Randomized-Block Designs

In three-level designs, the sampling involves clusters (such as schools) and subclusters (such as classrooms). In three-level randomized-block designs, both treatments are given to some individuals within every school. There are two variations of the three-level randomized-block experiment. One variant assigns subclusters (e.g., classrooms) to treatments, so that every individual within the same subcluster (classroom) receives the same treatment, but different subclusters within the same cluster (different classrooms within the same school) receive different treatments. The other variant assigns individuals within subclusters (classrooms) to treatments, so that some individuals receive each treatment in every classroom.

Suppose that the intervention effect at the population level is $(\mu_1 - \mu_2)$ in the units in which the outcome is measured (e.g., test score scale points). Statistical power again depends on the intervention effect via the effect size or standardized intervention effect (sometimes called Cohen's *d*)

$$\delta = \frac{\mu_1 - \mu_2}{\sigma},$$

where σ is the total population standard deviation of the outcome within-treatment groups. That is,

$$\sigma^2 = \sigma_S^2 + \sigma_C^2 + \sigma_W^2\,.$$

In three-level designs, recall that two indices—which were defined as the school-level (cluster) ICC ρ_S (defined in equation [10] above) and the classroom-level (subcluster) ICC ρ_C (defined in equation [11] above)—are necessary to characterize the intraclass correlation structure.

In two-level randomized-block designs, it was observed that power also depends on the degree to which treatment effects vary across clusters. This observation is also true in three-level randomized-block designs. In three-level randomized-block designs that assign treatments within subclusters (e.g., classrooms), power also depends on the degree to which treatment effects vary across subclusters. The treatment effect heterogeneity across clusters is characterized via the

ratio, ω_S, of variation of the treatment effects across clusters to the variation of untreated cluster means. Thus, ω_S can be characterized via

$$\omega_S = \sigma_{T\times S}^2/\sigma_S^2, \tag{24}$$

where $\sigma_{T\times S}^2$ is the treatment by cluster interaction and σ_S^2 is the variance of the cluster means. Similarly, in the case of randomized-block designs that assign treatments within subclusters, the treatment effect heterogeneity across subclusters is characterized via the ratio, ω_C, of variation of the treatment effects across subclusters to the variation of untreated subcluster means. Thus, ω_C can be characterized via

$$\omega_C = \sigma_{T\times C}^2/\sigma_C^2, \tag{25}$$

where $\sigma_{T\times C}^2$ is the treatment by subcluster interaction variance and σ_C^2 is the variance of the untreated subcluster means within clusters.

As in the case of two-level randomized-block designs, when the usual statistical model is used for power analysis, under very mild additional assumptions $\sigma_{T\times S}^2$ is less than σ_S^2, so that ω_S is usually less than one and can be as small as zero in cases where the treatment effect is very similar across clusters. In the same way, when the usual statistical model is used for power analysis, under very mild additional assumptions $\sigma_{T\times C}^2$ is less than σ_C^2, so that ω_C is usually less than one and can be as small as zero in cases where the treatment effect is very similar across subclusters.

Power computations for three-level randomized-block designs are quite similar to power computations for two-level randomized-block designs and two- and three-level hierarchical designs. In each case, power computation begins with computing a design effect and then using that design effect to compute the operational effect size. This operational effect size is then used along with the operational sample size to obtain the statistical power from tables of the power of the one-sample *t* test (such as Table 2). The computation of power for these designs is described in detail in Appendix C.

Chapter 9: Conclusions

This paper has provided an introduction to statistical power analysis in complex designs involving two- or three-stage cluster sampling. It shows how to use the concepts of operational effect sizes and sample sizes to compute statistical power using the power tables constructed for simple sampling designs. This formulation also explains how the clustering structure described by intraclass correlation coefficients (ICCs) influences operational effect size and, therefore, statistical power. Additionally, this paper details how the use of covariates can increase power by increasing the operational effect size.

Several general conclusions about design follow. The first is that in hierarchical designs, clustering will always decrease statistical power compared with a design that does not involve clustering. The difference in power will be determined by the design effect, which is a function of the experimental design and the ICC (e.g., of ρ). We note that the design effect we define differs from the usual design effect described in Kish (1965) and elsewhere. For example, in a two-level hierarchical design, the design effect would be $\sqrt{n}$ if there were no clustering, but it will never be larger than $\sqrt{1/\rho}$ in a design with clustering—no matter how large the within-cluster sample size (n) becomes. The design effect (and the power) approach the maximum when n is quite modest, so that increasing n beyond this point has little effect on power. In making decisions about allocation of sample size, it is therefore better (in terms of statistical power) to have a larger number of clusters than a larger number of individuals within clusters. The use of covariates can increase power substantially and, therefore, is always desirable.

Power for randomized-block designs may be computed in a fashion similar to hierarchical designs. When they are feasible, randomized-block designs will always have higher power than hierarchical designs for the same sample size. In randomized-block designs, the power depends on the heterogeneity of treatment effects across clusters (blocks), and this influence can be profound. Thus, the power advantage of randomized-block designs is most substantial when treatment effects are reasonably homogeneous across clusters. In fact, in randomized-block designs where treatment effects are homogeneous across clusters, power will generally be greater than the power in designs that do not involve clustering.

Appendix A: Design Effects in Two- or Three-Level Hierarchical Designs With and Without Covariates

No Covariates	With Covariates
Two-level hierarchical design	
$\sqrt{\frac{n}{1+(n-1)\rho}}$	$\sqrt{\frac{2m}{2m-q_S}}\sqrt{\frac{n}{1+(n-1)\rho-\left[R_W^2+\left(nR_S^2-R_W^2\right)\rho\right]}}$
Three-level hierarchical design	
$\sqrt{\frac{pn}{1+(pn-1)\rho_S+(n-1)\rho_C}}$	$\sqrt{\frac{2m}{2m-q_S}}\sqrt{\frac{pn}{1+(pn-1)\rho_S+(n-1)\rho_C-\left[R_W^2+(pnR_S^2-R_W^2)\rho_S+(nR_C^2-R_W^2)\rho_C\right]}}$

Note: See the text for definitions of symbols used in this table. They can be found on the following pages:

Design Effects in Two-Level Hierarchical Designs Without Covariates
Pages 17

Design Effects in Two-Level Hierarchical Designs With Covariates
Pages 20-21 and Formula 8

Design Effects in Three-Level Hierarchical Designs Without Covariates
Pages 25-26 and Formula 12

Design Effects in Three-Level Hierarchical Designs With Covariates
Pages 28-29 and Formula 14

Appendix B: Design Effects in Two- or Three-Level Randomized-Block Designs With and Without Covariates

No Covariates	With Covariates
Two-level randomized block design	
$\sqrt{\frac{n}{2\left[1+\left(n\omega-1\right)\rho\right]}}$	$\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{n/2}{1+\left(n\omega-1\right)\rho-\left[R_W^2+\left(n\omega R_{TS}^2-R_W^2\right)\rho\right]}}$
Three-level randomized block design assigning treatments to subclusters	
$\sqrt{\frac{pn}{2\left[1+\left(pn\omega_S-1\right)\rho_S+(n-1)\rho_C\right]}}$	$\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{pn/2}{1+\left(pn\omega_S-1\right)\rho_S+(n-1)\rho_C-\left[R_W^2+\left(pn\omega_S R_{TS}^2-R_W^2\right)\rho_S+\left(nR_C^2-R_W^2\right)\rho_C\right]}}$
Three-level randomized block design assigning treatments to individuals	
$\sqrt{\frac{pn}{2\left[1+\left(pn\omega_S-1\right)\rho_S+(n\omega_C-1)\rho_C\right]}}$	$\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{pn/2}{1+\left(pn\omega_S-1\right)\rho_S+(n\omega_C-1)\rho_C-\left[R_W^2+\left(pn\omega_S R_{TS}^2-R_W^2\right)\rho_S+\left(n\omega_C R_{TC}^2-R_W^2\right)\rho_C\right]}}$

Note: See the text for definitions of symbols used in this table. They can be found on the following pages:

Design Effects in Two-Level Randomized-Block Designs Without Covariates:
Pages 34-35 and Formula 18

Design Effects in Two-Level Randomized-Block Designs With Covariates
Pages 38-39 and Formula 22

Three-Level Randomized-Block Designs Assigning Treatments to Subclusters Without Covariates
Appendix C C-1 and Formula 26

Three-Level Randomized-Block Designs Assigning Treatments to Subclusters With Covariates
Appendix C C-2, C-3 and Formula 28

Three-Level Randomized-Block Designs Assigning Treatments to Individuals Without Covariates
Appendix C C-5 and Formula 31

Three-Level Randomized-Block Designs Assigning Treatments to Individuals With Covariates
Appendix C C-6 and Formulae 33 and 34

Appendix C: Computing Power in Three-Level Randomized-Block Designs

This appendix sketches the computation of statistical power in three-level randomized-block designs where the highest level units are considered clusters (random effects). Two variations are considered. One design assigns intact classrooms (subclusters) to treatments. The other variation assigns individuals within classrooms to treatments. In each case, designs with and without covariates are considered. However, the computations are quite similar in every case (see Konstantopoulos 2008a).

Three-Level Randomized-Block Designs That Assign Treatment to Subclusters With No Covariates

Suppose that one is planning a three-level randomized-block experiment that will use a total of m clusters (typically schools), each with $2p$ subclusters, and will assign p of these subclusters (classrooms) within each cluster (school) to a treatment condition and p of the subclusters (classrooms) within each cluster (school) to a control condition. Assume that n individuals are within each subcluster (classroom), so that mpn individuals are assigned to each treatment and the total sample size is $N = 2mpn$.

If the treatment is assigned to subclusters within each of the m clusters, one enters the power table with operational sample size $N^T = m$. The operational effect size is

$$\Delta^T = \delta \sqrt{\frac{pn}{2\left[1+\left(pn\omega_S - 1\right)\rho_S + (n-1)\rho_C\right]}}, \tag{26}$$

where δ is the effect size, ρ_S is the school level ICC, ρ_C is the classroom level ICC, ω_S is one half of the ratio of variance of the treatment effects across clusters to the total cluster level variance, p is the number of subclusters (classrooms) assigned to each treatment within each cluster (school), and n is the sample size in each cluster. Note that the design effect

$$\sqrt{\frac{pn}{2\left[1+\left(pn\omega_S - 1\right)\rho_S + (n-1)\rho_C\right]}}$$

can be larger than one, so the operational effect size Δ^T can be larger than the actual effect size δ. However, the operational sample size $N^T = m$ is smaller than the actual sample size mpn assigned to each treatment, so the power in the design with clustering usually will not be larger than in the design without clustering.

Using the operational effect size makes it possible to compute statistical power and sample size requirements for analyses based on clustered samples using tables and computer programs designed for the one-sample t test. For example, entering Table 2 on the row given by the operational sample size N^T, and finding the column corresponding to the operational effect size Δ^T, one can read the power value.

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral *t*-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (related to the operational effect size) called the noncentrality parameter—specifically,

$$\lambda = \sqrt{m}\Delta^T = \delta\sqrt{\frac{mpn}{2\left[1+\left(pn\omega_S-1\right)\rho_S+(n-1)\rho_C\right]}}. \qquad (27)$$

Power is computed using equation (20) for a one-tailed test or (21) for a two-tailed test, as above, except that the noncentrality parameter (27) is used with $m - 1$ degrees of freedom.

Example. Return to the design considered earlier for a study to evaluate the effects of a second-grade supplemental reading intervention. Remember that the assumptions made here are intended for illustration only and are not intended to suggest values that will always be appropriate for other research studies. Continue to assume that a school-level ICC of about $\rho_S = 0.20$ is plausible and that a classroom-level ICC of about $\rho_c = 0.13$ is plausible. Assume that $n = 10$ students from each of $p = 2$ classrooms would receive each experimental condition from each school. Previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$ and that effects are likely to be fairly consistent across schools, so that a value of $\omega_S = 0.5$ is plausible (even fairly conservative). The initial plan was for a study that would include $m = 30$ schools. To determine the statistical power of this design, one first computes the operational effect size, using equation (26):

$$\Delta^T = (0.35)\sqrt{\frac{2(10)/2}{1+[2(10)(0.5)-1](0.20)+\left(10-1\right)(0.13)}} = (0.35)(1.587) = 0.555.$$

Entering Table 2 on the row corresponding to $N^T = m = 30$, one sees that the power for $\Delta^T = 0.56$ will be between 0.75 (the power for $\Delta^T = 0.50$) and 0.89 (the power for $\Delta^T = 0.60$). Interpolating between these two values (0.56 is six-tenths of the way from 0.5 to 0.6, so go six-tenths of the way between 0.75 and 0.89), one obtains a power level of 0.83.

Three-Level Randomized-Block Designs That Assign Treatment to Subclusters With Covariates

Now suppose that the analysis has q_S ($0 \le q_S < m - 1$) cluster-level covariates, q_C ($0 \le q_C < mp - q_S - 1$) subcluster-level covariates, and q_W ($0 \le q_W < N - q_S - q_C - 2$) individual-level covariates[4]. For example, a design with $q_W = 1$, $q_C = 1$, and $q_S = 1$ might arise if a pretest were used (centered on subcluster means) as an individual-level covariate, subcluster means (centered on cluster means) on the pretest were used as a subcluster-level covariate, and cluster means on the pretest were used as a cluster-level covariate.

[4] Note that the possibility of having 0 (no) covariates at a given level has been included in the previous section, and that adding covariates necessitates modifications of the operational sample size and the operational effect sizes.

When covariates are used in the design, both the operational effect size and the operational sample size are slightly modified. As was the case in the two-level randomized-block design, one would enter Table 2 with operational sample size $N_A^T = m - q_s$.

The operational effect size is increased to an extent that depends on how much the covariates explain between-cluster variance in treatment effects, variance between subcluster means, and variance within subclusters. R_W^2 is the amount of within-subcluster variance the covariates explain, R_{TS}^2 is the amount of between-cluster (school) variance in treatment effects the covariates explain, and R_C^2 is the amount of between-subcluster (classroom) variance the covariates explain. One can think of R_W^2, R_C^2, and R_{TS}^2 as proportions of variance accounted for (squared correlations) in the usual way. The covariate-adjusted operational effect size is

$$\Delta_A^T = \delta\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{pn/2}{1+(pn\omega_S-1)\rho_S+(n-1)\rho_C-\left[R_W^2+\left(pn\omega_S R_{TS}^2-R_W^2\right)\rho_S+\left(nR_C^2-R_W^2\right)\rho_C\right]}}. \quad (28)$$

The covariate-adjusted operational effect size is generally larger than the unadjusted design effect Δ^T (and cannot be smaller than Δ^T) because R_{TS}^2, R_C^2, and R_W^2 are generally larger than zero. Hence, the term in square brackets in the denominator of the second term of the design effect

$$\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{pn/2}{1+(pn\omega_S-1)\rho_S+(n-1)\rho_C-\left[R_W^2+\left(pn\omega_S R_{TS}^2-R_W^2\right)\rho_S+\left(nR_C^2-R_W^2\right)\rho_C\right]}}$$

is generally positive. In computing Δ_A^T, it may be more convenient to break the denominator of the second term of the design effect into two parts. One part,

$$A = 1 + (pn\omega_S - 1)\rho_S + (n - 1)\rho_C,$$

reflects the part due to clustering, and the second part,

$$B = R_W^2 + (pn\omega_S R_{TS}^2 - R_W^2)\rho_S + (nR_C^2 - R_W^2)\,\rho_C,$$

reflects the adjustment for the effects of covariates, so that

$$\Delta_A^T = \delta\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{pn/2}{A-B}}. \quad (29)$$

Using the operational effect size makes it possible to compute statistical power and sample-size requirements for analyses based on clustered samples using these tables and computer programs designed for the one-sample *t* test. For example, entering Table 2 on the row given by the operational sample size N_A^T, and finding the column corresponding to the operational effect size Δ_A^T, one can read the power value.

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral *t*-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (related to the operational effect size) called the noncentrality parameter—specifically,

$$\lambda_A = \sqrt{m - q_S}\,\Delta_A^T = \delta\sqrt{\frac{mpn/2}{1+(pn\omega_S - 1)\rho_S + (n-1)\rho_C - \left[R_W^2 + (pn\omega_S R_{TS}^2 - R_W^2)\rho_S + (nR_C^2 - R_W^2)\rho_C\right]}}. \quad (30)$$

Power is computed using equation (20) for a one-tailed test or (21) for a two-tailed test, as above, except that the noncentrality parameter (30) is used with $m - q_s - 1$ degrees of freedom.

Example. Return to the design considered earlier for a study to evaluate the effects of a second-grade supplemental reading intervention. Continue to assume that a school-level ICC of about $\rho_S = 0.20$ is plausible and that a classroom-level ICC of about $\rho_C = 0.13$ is plausible. Assume that $n = 10$ students from each of $p = 2$ classrooms would participate from each school. Previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$ and that effects are likely to be fairly consistent across schools, so that a value of $\omega_S = 0.5$ is plausible (even fairly conservative). Suppose that pretreatment reading test scores are available and that classroom-centered individual pretest scores, school-centered classroom mean pretest scores, and school mean pretest scores will be used as covariates. Thus, $q_W = q_C = q_S = 1$. There is some evidence that values of $R_W^2 = 0.5$, $R_C^2 = 0.6$, and $R_S^2 = 0.8$ are plausible. We again assume that about half of the total variance between schools explained by the covariates is due to the ability of the covariates to predict treatment effect variability, so that $R_{TS}^2 = 0.4$. We clarify that these assumptions are intended for illustration only and are not intended to suggest values that will always be appropriate for other research studies. The initial plan was for a study that would include $m = 20$ schools.

To determine the statistical power of this design, one first computes the operational effect size via equation (29), using

$$A = 1 + [(2)(10)(0.5) - 1](0.2) + (10 - 1)(0.13) = 3.970$$

and

$$B = 0.5 + [(2)(10)(0.5)(0.4) - 0.5](0.2) + [(10)(0.6) - 0.5](0.13) = 1.915$$

to obtain

$$\Delta_A^T = \sqrt{0.35}\sqrt{\frac{20 \quad 2(10)/2}{19 \quad 3.970 - 1.915}} = (0.35)(1.026)(2.205) = 0.79$$

Entering Table 2 on the row corresponding to $N_A^T = m - q_s = 19$, one sees that the power for $\Delta^T = 0.79$ is slightly less than 0.91 (the power value listed for $\Delta^T = 0.80$).

One might regard this high statistical power as providing a margin of safety in case any assumptions are somewhat optimistic. Alternatively, because power of 0.91 may be higher than necessary, one might consider altering the design to decrease costs while maintaining acceptable statistical power. For example, one might consider decreasing the number of schools to $m = 15$. That decrease would have little effect on the operational effect size (it would now be 0.80), but reading power values from the row of the table for $N^T = m - 1 = 14$, one sees that the power is 0.79.

Three-Level Randomized-Block Designs That Assign Treatment Within Subclusters With No Covariates

Suppose that one is planning a three-level randomized-block experiment that will use a total of m clusters (typically schools), each with p subclusters (classrooms) and $2n$ individuals within each of these subclusters. The design will assign n individuals within each of these subclusters (classrooms) to a treatment condition and n individuals within each of these subclusters (classrooms) to a control condition. Thus, mpn individuals are assigned to each treatment, and the total sample size is $N = 2mpn$.

If the actual number of clusters in the experiment is m, one enters the power table with operational sample size $N^T = m$. The operational effect size is

$$\Delta^T = \delta \sqrt{\frac{pn}{2\left[1 + \left(pn\omega_S - 1\right)\rho_S + (n\omega_C - 1)\rho_C\right]}}, \tag{31}$$

where δ is the effect size, ρ_S is the school level ICC, ρ_C is the classroom level ICC, ω_S is one half of the ratio of variance of the treatment effects across clusters to the total cluster level variance, ω_C is one half of the ratio of variance of the treatment effects across subclusters to the total subcluster level variance, p is the number of subclusters (classrooms) within each cluster (school), and n is the number of individuals assigned to each treatment within each subcluster. Note that the design effect

$$\Delta^T = \delta \sqrt{\frac{pn}{2\left[1 + \left(pn\omega_S - 1\right)\rho_S + (n\omega_C - 1)\rho_C\right]}}$$

can be larger than one, so the operational effect size Δ^T can be larger than the actual effect size δ. However, the operational sample size $N^T = m$ is smaller than the actual sample size mpn assigned to each treatment, so the power in the design with clustering usually will not be larger than in the design without clustering.

Using the operational effect size makes it possible to compute statistical power and sample size requirements for analyses based on clustered samples using tables and computer

programs designed for the one-sample *t* test. For example, entering Table 2 on the row given by the operational sample size N^T, and finding the column corresponding to the operational effect size Δ^T, one can read the power value.

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral *t*-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (related to the operational effect size) called the noncentrality parameter—specifically,

$$\lambda = \sqrt{m}\Delta^T = \delta\sqrt{\frac{mpn}{2\left[1+\left(pn\omega_S-1\right)\rho_S+(n\omega_C-1)\rho_C\right]}}. \qquad (32)$$

Power is computed using equation (20) for a one-tailed test or (21) for a two-tailed test, as above.

Example. Return to the design considered earlier for a study to evaluate the effects of a second-grade supplemental reading intervention. Continue to assume that a school-level ICC of about $\rho_S = 0.20$ is plausible and that a classroom-level ICC of about $\rho_C = 0.13$ is plausible. Assume that $n = 10$ students from each of $p = 2$ classrooms from each school would be assigned to each treatment within each classroom. Previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$ and that effects are likely to be fairly consistent across schools, so that a value of $\omega_S = \omega_C = 0.5$ is plausible (even fairly conservative). The initial plan was for a study that would include $m = 30$ schools. We again clarify that these assumptions are intended for illustration only and are not intended to suggest values that will always be appropriate for other research studies. To determine the statistical power of this design, one first computes the operational effect size, using equation (31):

$$\Delta^T = (0.35)\sqrt{\frac{2(10)}{2\left\{1+\left[2(10)(0.5)-1\right](0.20)+\left[(10)(0.5)-1\right](0.13)\right\}}} = (0.35)(1.736) = 0.607$$

Entering Table 2 on the row corresponding to $N^T = m = 30$, one sees that the power for

$\Delta^T = 0.61$ will be between 0.89 (the power for $\Delta^T = 0.60$) and 0.96 (the power for $\Delta^T = 0.70$). Interpolating between these two values (0.61 is one-tenth of the way from 0.6 to 0.7, so go one-tenth of the way between 0.89 and 0.96), one obtains a power level of 0.90.

Three-Level Randomized-Block Designs That Assign Treatment Within Subclusters With Covariates

Now suppose that the analysis has q_S ($0 \le q_S < m - 1$) cluster-level covariates, q_C ($0 \le q_C < mp - q_S - 1$) subcluster-level covariates, and q_W ($0 \le q_W < N - q_S - q_C - 2$) individual-level covariates.[5] For example, a design with $q_W = 1$, $q_C = 1$, and $q_S = 1$ might arise if a pretest were used (centered on subcluster means) as an individual-level covariate, subcluster means (centered on cluster means) on the pretest were used as a cluster-level covariate, and cluster means on the pretest were used as a covariate.

When covariates are used in the design, the operational effect size is modified (typically increased) to an extent that depends on how much the covariates explain between-cluster and between-subcluster variance in treatment effects and within-cluster variance. R_W^2 is the amount of within-cluster variance the covariates explain, R_{TS}^2 is the amount of between-cluster (between-schools) variance in treatment effects the covariates explain, and R_{TC}^2 is the amount of between-subcluster but within-cluster (between-classroom within-schools) variance in treatment effects the covariates explain. One can think of R_W^2, R_{TC}^2, and R_{TS}^2 as proportions of variance accounted for (squared correlations) in the usual way. The covariate-adjusted operational effect size is

$$\Delta_A^T = \delta \sqrt{\frac{m}{m - q_S}} \sqrt{\frac{pn/2}{1 + (pn\omega_S - 1)\rho_S + (n\omega_C - 1)\rho_C - \left[R_W^2 + (pn\omega_S R_{TS}^2 - R_W^2)\rho_S + (n\omega_C R_{TC}^2 - R_W^2)\rho_C\right]}}. \quad (33)$$

The covariate-adjusted operational effect size is generally larger than the unadjusted design effect Δ^T (and cannot be smaller than Δ^T) because R_{TS}^2, R_{TC}^2, and R_W^2 are generally larger than zero. Hence, the term in square brackets in the denominator of the second term of the design effect

$$\sqrt{\frac{m}{m - q_S}} \sqrt{\frac{pn/2}{1 + (pn\omega_S - 1)\rho_S + (n\omega_C - 1)\rho_C - \left[R_W^2 + (pn\omega_S R_{TS}^2 - R_W^2)\rho_S + (n\omega_C R_{TC}^2 - R_W^2)\rho_C\right]}}$$

is generally positive.

In computing Δ_A^T, it may be more convenient to break the denominator of the design effect into two parts. One part,

$$A = 1 + (pn\omega_S - 1)\rho_S + (n\omega_C - 1)\rho_C,$$

reflects the impact of clustering, and the second part,

[5] Note that the possibility of having 0 (no) covariates at a given level has been included in the previous section, and that adding covariates necessitates modifications of the operational sample size and the operational effect sizes.

$$B = R_W^2 + (pn\omega_S R_{TS}^2 - R_W^2)\rho_S + (n\omega_C R_{TC}^2 - R_W^2)\,\rho_C,$$

reflects the adjustment for the effects of covariates, so that

$$\Delta_A^T = \delta\sqrt{\frac{m}{m-q_S}}\sqrt{\frac{pn/2}{A-B}}\,. \tag{34}$$

Using the operational effect size makes it possible to compute statistical power and sample-size requirements for analyses based on clustered samples using power tables designed for the one-sample *t* test. For example, entering Table 2 on the row given by the operational sample size N_A^T, and finding the column corresponding to the operational effect size Δ_A^T, one can read the power value.

An alternative method for computing statistical power in clustered designs is to use a computer program that has a built-in function that computes the noncentral *t*-distribution. To use such a function to compute the statistical power of a test, one must provide the program with a value of an index λ (related to the operational effect size) called the noncentrality parameter—specifically,

$$\begin{aligned}\lambda_A &= \sqrt{m-q_s}\,\Delta_A^T \\ &= \delta\sqrt{\frac{mpn/2}{1+\left(pn\omega_S-1\right)\rho_S+(n\omega_C-1)\rho_C-\left[R_W^2+\left(pn\omega_S R_{TS}^2-R_W^2\right)\rho_S+\left(n\omega_C R_{TC}^2-R_W^2\right)\rho_C\right]}}\,.\end{aligned} \tag{35}$$

Power is computed using equation (20) for a one-tailed test or (21) for a two-tailed test, as above, except that the covariate-adjusted noncentrality parameter (35) is used with $m - 1 - q_S$ degrees of freedom.

Example. Returning to the design considered earlier for a study to evaluate the effects of a second-grade supplemental reading intervention, continue to assume that a school-level ICC of about $\rho_S = 0.20$ is plausible and that a classroom-level ICC of about $\rho_C = 0.13$ is plausible. Assume that $n = 10$ students would be assigned to each treatment from each of $p = 2$ classrooms in each school. Previous studies of this intervention suggest that the effect size is likely to be $\delta = 0.35$ and that effects are likely to be fairly consistent across schools so that a value of $\omega_S = \omega_C = 0.5$ is plausible (even fairly conservative). Suppose that pretreatment reading test scores are available and that classroom-centered individual pretest scores, school-centered classroom mean pretest scores, and school mean pretest scores will be used as covariates. Thus, $q_W = q_C = q_S = 1$. There is some evidence that values of $R_W^2 = 0.5$, $R_{TC}^2 = 0.6$, and $R_{TS}^2 = 0.8$ are plausible. We again assume that about half of the total variance between schools explained by the covariates is due to the ability of the covariates to predict treatment effect variability, so that $R_{TS}^2 = 0.4$. We also assume that about half of the total variance between classrooms within schools explained by the covariates is due to the ability of the covariates to predict treatment effect variability, so that $R_{TC}^2 = 0.3$. However, we clarify that these assumptions are intended for illustration only and are not intended to suggest values that

will always be appropriate for other research studies. The initial plan was for a study that would include $m = 15$ schools.

To determine the statistical power of this design, one first computes the operational effect size via equation (34), using

$$A = 1 + [(2)(10)(0.5) - 1](0.2) + [(10)(0.5) - 1](0.13) = 3.320$$

and

$$B = 0.5 + [(2)(10)(0.5)(0.4) - 0.5](0.2) + [(10)(0.5)(0.3) - 0.5](0.13) = 1.33$$

to obtain

$$\Delta_A^T = 0.35\sqrt{\frac{15}{14}}\sqrt{\frac{2(10)/2}{3.320 - 1.32}} = (0.35)(1.035)(2.24) = 0.812 .$$

Entering Table 2 on the row corresponding to $N_A^T = m - 1 = 14$, one sees that the power for $\Delta_A^T = 0.81$ is slightly more than that listed for $\Delta_A^T = 0.80$, which is 0.79.

Appendix D: Multilevel Models Defining Tests for Treatment Effects

This appendix describes the multilevel models on which the power computations are based. When designs are balanced, the power calculations are exact because exact tests for the treatment effect based on the analysis of variance are possible.

Two-Level Hierarchical Designs With No Covariates.

Suppose that m clusters of size n are assigned to each treatment. Let Y_{ij} be the j^{th} observation in the i^{th} cluster. Then the level 1 (individual-level) model is

$$Y_{ij} = \beta_{0i} + \varepsilon_{ij} \quad , i = 1, \ldots, 2m; j = 1, \ldots n,$$

where β_{0i} is the mean of the i^{th} cluster and the ε_{ij} are independently normally distributed with mean 0 and variance σ_W^2. The level 2 (cluster-level) model is

$$\beta_{0i} = \gamma_{00} + \gamma_{01} T_i + \eta_{0i}, i = 1, ..., 2m,$$

where γ_{00} is the grand mean, γ_{01} is the treatment effect, T_i is a treatment indicator coded $T_i = ½$ for treatment clusters and $T_i = -½$ for the control clusters, and the η_{0i} are independently normally distributed with mean 0 and variance σ_S^2.

The treatment effect size δ is defined as

$$\delta = \frac{\gamma_{01}}{\sqrt{\sigma_S^2 + \sigma_W^2}}.$$

The intraclass correlation coefficient (ICC) is defined in terms of the variances as

$$\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\gamma_{01} = 0$.

Two-Level Hierarchical Designs With Covariates

Suppose that m clusters of size n are assigned to each treatment. Let Y_{ij} be the j^{th} observation in the i^{th} cluster. Now suppose that q covariates are modeled at the individual level and r covariates are modeled at the cluster level. Thus, the level 1 (individual-level) model is

$$Y_{ij} = \beta_{0i} + \beta_1 X_{1ij} + \cdots + \beta_q X_{qij} + \varepsilon_{ij} \quad , i = 1, \ldots, 2m; j = 1, \ldots n,$$

where β_{0i} is the covariate-adjusted mean of the i^{th} cluster, $\beta_1, \ldots, \beta_q$ are the (fixed) effects of the individual-level covariate, $X_{1ij}, \ldots, X_{qij}$ are the values of the individual-level covariates

(centered on cluster means), and the ε_{ij} are independently normally distributed with mean 0 and variance σ_W^2. The level 2 (cluster-level) model is

$$\beta_{0i} = \gamma_{00} + \gamma_{01}T_i + \gamma_{02}W_{1i} + \ldots + \gamma_{0(r+1)}W_{ri} + \eta_{0i},\ i = 1, ..., 2m,$$

where γ_{00} is the covariate-adjusted grand mean, γ_{01} is the treatment effect, $\gamma_{02}, \ldots, \gamma_{0(r+1)}$ are the effects of cluster-level covariates, $W_{1i}, \ldots, W_{ri}$ are the values of the cluster-level covariates for cluster i, T_i is a treatment indicator coded $T_i = ½$ for treatment clusters and $T_i = -½$ for the control clusters, and the η_{0i} are independently normally distributed with mean 0 and variance σ_{AS}^2. Note that covariates are treated as having fixed effects.

In this model, the treatment effect size δ is still defined in terms of the unadjusted total standard deviation—that is,

$$\delta = \frac{\gamma_{01}}{\sqrt{\sigma_S^2 + \sigma_W^2}}.$$

The ICC is also defined in terms of the unadjusted variances as

$$\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2},$$

and the covariate outcome correlations are defined in terms of the adjusted and unadjusted variances as

$$R_S^2 = 1 - \frac{\sigma_{AS}^2}{\sigma_S^2}$$

and

$$R_W^2 = 1 - \frac{\sigma_{AW}^2}{\sigma_W^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\gamma_{01} = 0$.

Three-Level Hierarchical Designs With No Covariates

Suppose that m clusters, each with p subclusters of size n, are assigned to each treatment. Let Y_{ijk} be the k^{th} observation in j^{th} subcluster of the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ijk} = \beta_{0ij} + \varepsilon_{ijk}\ ,\ i = 1, \ldots, 2m; j = 1, \ldots p; k = 1, \ldots n,$$

where β_{0ij} is the mean of the j^{th} subcluster in the i^{th} cluster, and the ε_{ijk} are independently normally distributed with mean 0 and variance σ_W^2. The level 2 (subcluster-level) model is

$$\beta_{0ij} = \gamma_{00i} + \eta_{0ij}, j = 1, ..., p, i = 1, ..., 2m,$$

where γ_{00} is the mean of the i^{th} cluster and the η_{0ij} are independently normally distributed with mean 0 and variance σ_C^2. The level 3 (cluster-level) model is

$$\gamma_{00i} = \pi_{00} + \pi_{01} T_i + \xi_{0i}, i = 1, ..., 2m,$$

where π_{00} is the grand mean, π_{01} is the treatment effect, T_i is a treatment indicator coded $T_i = ½$ for treatment clusters and $T_i = -½$ for the control clusters, and the ξ_{0i} are independently normally distributed with mean 0 and variance σ_S^2.

The treatment effect size δ is defined as

$$\delta = \frac{\pi_{01}}{\sqrt{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}}.$$

The ICCs are defined in terms of the variances as

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}$$

and

$$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\pi_{01} = 0$.

Three-Level Hierarchical Designs With Covariates

Suppose that m clusters, each with p subclusters of size n, are assigned to each treatment. Now suppose that $q \geq 0$ covariates are at the cluster level, $r \geq 0$ covariates are at the subcluster level, and $s \geq 0$ covariates are at the individual level. Let Y_{ijk} be the k^{th} observation in the j^{th} subcluster of the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ijk} = \beta_{0ij} + \beta_1 X_{1ijk} + \cdots + \beta_q X_{qijk} + \varepsilon_{ijk} \quad , i = 1, \ldots, 2m; j = 1, \ldots p; k = 1, \ldots n,$$

where β_{0ij} is the covariate-adjusted mean of the j^{th} subcluster in the i^{th} cluster, $\beta_1, \ldots, \beta_q$ are the effects of the individual-level covariates (which are fixed effects), $X_{1ijk}, \ldots, X_{qijk}$ are the values of the individual-level covariates (centered on subcluster means), and the ε_{ijk} are

independently normally distributed with mean 0 and variance σ_{AW}^2. The level 2 (subcluster-level) model is

$$\beta_{0ij} = \gamma_{00i} + \gamma_1 Z_{1ij} + \cdots + \gamma_r Z_{rij} + \eta_{0ij}, j = 1, ..., p, i = 1, ..., 2m,$$

where γ_{00i} is the covariate-adjusted mean of the i^{th} cluster, $\gamma_1, \ldots, \gamma_r$ are the effects of the level 2 covariates (which are fixed effects), $Z_{1ij}, \ldots, Z_{rij}$ are the values of the subcluster-level covariates (centered on cluster means), and the η_{0ij} are independently normally distributed with mean 0 and variance σ_{AC}^2. The level 3 (cluster-level) model is

$$\gamma_{00i} = \pi_{00} + \pi_{01}T_i + \pi_2 W_{1i} + \ldots + \pi_{s+1}W_{si} + \xi_{0i}, i = 1, ..., 2m,$$

where π_{00} is the covariate-adjusted grand mean, π_{01} is the treatment effect, $\pi_{02}, \ldots, \pi_{0(s+1)}$ are the effects of the level 3 covariates, T_i is a treatment indicator coded $T_i = ½$ for treatment clusters and $T_i = -½$ for the control clusters, $\pi_1, \ldots, \pi_{s+1}$ are effects of the covariates, $W_{1i}, \ldots, W_{si}$ are the values of the cluster-level covariate, and the ξ_{0i} are independently normally distributed with mean 0 and variance σ_{AS}^2.

The treatment effect size δ is defined in terms of the unadjusted variances—that is,

$$\delta = \frac{\pi_{01}}{\sqrt{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}}.$$

The ICCs are defined in terms of the unadjusted variances as

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2},$$

and

$$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}.$$

The covariates outcome correlations are defined in terms of the adjusted and unadjusted variances as

$$R_S^2 = 1 - \frac{\sigma_{AS}^2}{\sigma_S^2},$$

and

$$R_C^2 = 1 - \frac{\sigma_{AC}^2}{\sigma_C^2}$$

$$R_W^2 = 1 - \frac{\sigma_{AW}^2}{\sigma_W^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\pi_{01} = 0$.

Two-Level Randomized-Block Designs With No Covariates

Suppose that there are m clusters of size $2n$ and that n individuals in each cluster are assigned to each treatment. Let Y_{ij} be the j^{th} observation in the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ij} = \beta_{0i} + \beta_{1i}T_i + \varepsilon_{ij} \quad , i = 1, \ldots, m; j = 1, \ldots, 2n,$$

where β_{0i} is the mean of the i^{th} cluster, T_i is a treatment indicator coded $T_i = ½$ for individuals assigned to treatment and $T_i = -½$ for the individuals assigned to control, and the ε_{ij} are independently normally distributed with mean 0 and variance σ_W^2. The level 2 (cluster-level) model is

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, i = 1, ..., m$$

and

$$\beta_{1i} = \gamma_{10} + \eta_{1i}, i = 1, ..., m,$$

where γ_{00} is the grand mean, γ_{10} is the (mean) treatment effect, the η_{0i} are independently normally distributed with mean 0 and variance σ_{Bl}^2, and the η_{1i} are independently normally distributed with mean 0 and variance $2\sigma_{T\times S}^2$.

The treatment effect size δ is defined as

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma_S^2 + \sigma_W^2}}, \text{ where } \sigma_S^2 = \sigma_{Bl}^2 + \sigma_{T\times S}^2.$$

The ICC is defined in terms of the variances as $\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2}$,

and the heterogeneity parameter ω is defined as

$$\omega = \frac{\sigma_{T\times S}^2}{\sigma_S^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\gamma_{10} = 0$.

Two-Level Randomized-Block Design With Covariates

Suppose that there are m clusters of size $2n$ and that n individuals in each cluster are assigned to each treatment. Now suppose that there are q covariates at the individual level and s covariates at the cluster level. Let Y_{ij} be the j^{th} observation in the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ij} = \beta_{0i} + \beta_{1i}T_{ij} + \beta_1 X_{1ij} + \cdots + \beta_q X_{qij} + \varepsilon_{ij} \quad , i = 1, \ldots, m; j = 1, \ldots, 2n,$$

where β_{0i} is the covariate-adjusted mean of the i^{th} cluster, T_{ij} is a treatment indicator coded $T_{ij} = ½$ for individuals assigned to treatment and $T_{ij} = -½$ for the individuals assigned to control, $\beta_1, \ldots, \beta_q$ are the effects of the individual-level covariates (which are fixed effects), and the ε_{ij} are independently normally distributed with mean 0 and variance σ_{AW}^2. The level 2 (cluster-level) model is

$$\beta_{0i} = \gamma_{00} + \gamma_{02}W_{1i} + \ldots + \gamma_{0(r+1)}W_{ri} + \eta_{0i}, i = 1, ..., m,$$

and

$$\beta_{1i} = \gamma_{10} + \gamma_{12}W_{1i} + \ldots + \gamma_{1(r+1)}W_{ri} + \eta_{1i}, i = 1, ..., m,$$

where γ_{00} is the covariate-adjusted grand mean, γ_{10} is the covariate-adjusted mean treatment effect, $\gamma_{02}, \ldots, \gamma_{0(r+1)}$ are the effects of the level 2 covariates on the mean, $\gamma_{02}, \ldots, \gamma_{0(r+1)}$ are the effects of the level 2 covariates on the cluster-specific treatment effects, the η_{0i} are independently normally distributed with mean 0 and variance σ_{ABI}^2, and the η_{1i} are independently normally distributed with mean 0 and variance $2\sigma_{ATxC}^2$.

The treatment effect size δ is defined as

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma_S^2 + \sigma_W^2}}.$$

The ICC is defined in terms of the unadjusted variances as $\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2}$,

and the heterogeneity parameter ω is defined in terms of the unadjusted variances as

$$\omega = \frac{\sigma_{T\times S}^2}{\sigma_S^2}.$$

The covariate outcome correlations are defined in terms of the adjusted and unadjusted variances as

$$R^2_{T\times S} = 1 - \frac{\sigma^2_{AT\times S}}{\sigma^2_{T\times S}}$$

and

$$R^2_W = 1 - \frac{\sigma^2_{AW}}{\sigma^2_W}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\gamma_{10} = 0$.

Three-Level Randomized-Block Designs Assigning Subclusters With No Covariates

Suppose that there are m clusters, each with $2p$ subclusters of size n, and that half of the subclusters in each cluster are assigned to each treatment. Let Y_{ijk} be the k^{th} observation in j^{th} subcluster of the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ijk} = \beta_{0ij} + \varepsilon_{ijk} \quad , i = 1, \ldots, m; j = 1, \ldots, 2p; k = 1, \ldots n,$$

where β_{0ij} is the mean of the j^{th} subcluster in the i^{th} cluster and the ε_{ijk} are independently normally distributed with mean 0 and variance σ_W^2. The level 2 (subcluster-level) model is

$$\beta_{0ij} = \gamma_{00i} + \gamma_{01i}T_{ij} + \eta_{0ij}, j = 1, ..., 2p, i = 1, ..., m,$$

where γ_{00i} is the mean of the i^{th} cluster, γ_{01i} is the treatment effect in the i^{th} cluster, T_{ij} is a treatment indicator coded $T_{ij} = ½$ for treatment subclusters and $T_{ij} = -½$ for the control subclusters, and the η_{0ij} are independently normally distributed with mean 0 and variance σ_C^2.

The level 3 (cluster-level) model is

$$\gamma_{00i} = \pi_{00} + \xi_{0i}, i = 1, ..., m$$

and

$$\gamma_{01i} = \pi_{10} + \xi_{1i}, i = 1, ..., m,$$

where π_{00} is the grand mean, π_{10} is the average treatment effect, the ξ_{0i} are independently normally distributed with mean 0 and variance σ_{SBl}^2, and the ξ_{1i} are independently normally distributed with mean 0 and variance $2\sigma_{T\times S}^2$.

The treatment effect size δ is defined in terms of the unadjusted variances—that is,

$$\delta = \frac{\pi_{10}}{\sqrt{\sigma^2_S + \sigma^2_C + \sigma^2_W}}, \text{ where } \sigma^2_S = \sigma_{SBl}^2 + \sigma_{T\times S}^2.$$

The ICCs are defined in terms of the variances as

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}$$

and $$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2},$$

and the heterogeneity parameter ω_S is defined in terms of the variances as

$$\omega_S = \frac{\sigma_{T\times S}^2}{\sigma_S^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\pi_{10} = 0$.

Three-Level Randomized-Block Designs Assigning Subclusters With Covariates

Suppose that there are m clusters, each with $2p$ subclusters of size n, and that half of the subclusters in each cluster are assigned to each treatment. Now suppose that there are also $q \geq 0$ covariates at the cluster level, $r \geq 0$ covariates at the subcluster level, and $s \geq 0$ covariates at the individual level. Let Y_{ijk} be the k^{th} observation in j^{th} subcluster of the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ijk} = \beta_{0ij} + \beta_1 X_{1ijk} + \cdots + \beta_q X_{qijk} + \varepsilon_{ijk} \qquad , i = 1, \ldots, m; j = 1, \ldots, 2p; k = 1, \ldots n,$$

where β_{0ij} is the covariate-adjusted mean of the j^{th} subcluster in the i^{th} cluster, $\beta_1, \ldots, \beta_q$ are the effects of the individual-level covariates (which are fixed effects), $X_{1ijk}, \ldots, X_{qijk}$ are the values of the individual-level covariates (centered on subcluster means), and the ε_{ijk} are independently normally distributed with mean 0 and variance σ_{AW}^2. The level 2 (subcluster-level) model is

$$\beta_{0ij} = \gamma_{00i} + \gamma_{01i} T_{ij} + \gamma_2 Z_{1ij} + \cdots + \gamma_{r+1} Z_{rij} + \eta_{0ij}, j = 1, \ldots, 2p, i = 1, \ldots, m,$$

where γ_{00i} is the covariate-adjusted mean of the i^{th} cluster, γ_{01i} is the treatment effect in the i^{th} cluster, T_{ij} is a treatment indicator coded $T_{ij} = ½$ for treatment subclusters and $T_{ij} = -½$ for the control subclusters, $\gamma_2, \ldots, \gamma_{r+1}$ are the effects of the level 2 covariates (which are fixed effects), $Z_{1ij}, \ldots, Z_{rij}$ are the values of the subcluster-level covariates (centered on cluster means), and the η_{0ij} are independently normally distributed with mean 0 and variance σ_{AC}^2. The level 3 (cluster-level) model is

$$\gamma_{00i} = \pi_{00} + \pi_{01} W_{1i} + \ldots + \pi_{0s} W_{si} + \xi_{0i}, i = 1, \ldots, m,$$

and

$$\gamma_{01i} = \pi_{10} + \pi_{11} W_{1i} + \ldots + \pi_{1s} W_{si} + \xi_{1i}, i = 1, \ldots, m,$$

where π_{00} is the covariate-adjusted grand mean, π_{10} is the covariate-adjusted average treatment effect, π_{01}, …, π_{0s} are the effects of the level 3 covariates on the cluster means, π_{11}, …, π_{1s} are the effects of the level 3 covariates on the cluster-specific treatment effects, W_{1i}, …, W_{si} are the values of the cluster-level covariates, the ξ_{0i} are independently normally distributed with mean 0 and variance σ_{ASBl}^2, and the ξ_{1i} are independently normally distributed with mean 0 and variance $2\sigma_{ATxS}^2$.

The treatment effect size δ is defined in terms of the unadjusted variances—that is,

$$\delta = \frac{\pi_{10}}{\sqrt{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}}.$$

The ICCs are defined in terms of the unadjusted variances as

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}$$

and

$$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2},$$

and the heterogeneity parameter ω_S is defined in terms of the unadjusted variances as

$$\omega_S = \frac{\sigma_{T\times S}^2}{\sigma_S^2}.$$

The covariate outcome correlations are defined in terms of the adjusted and unadjusted variances as

$$R_{T\times S}^2 = 1 - \frac{\sigma_{AT\times S}^2}{\sigma_{T\times S}^2}$$

$$R_C^2 = 1 - \frac{\sigma_{AC}^2}{\sigma_C^2},$$

and

$$R_W^2 = 1 - \frac{\sigma_{AW}^2}{\sigma_W^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\pi_{10} = 0$.

Three-Level Randomized-Block Designs: Assigning Individuals Within Subclusters With No Covariates

Suppose that there are m clusters, each with p subclusters of size $2n$, and that half of the individuals in each subcluster are assigned to each treatment. Let Y_{ijk} be the k^{th} observation in the j^{th} subcluster of the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ijk} = \beta_{0ij} + \beta_{1ij}T_{ijk} + \varepsilon_{ijk}\ , i = 1, \ldots, m; j = 1, \ldots p; i = 1, \ldots 2n,$$

where β_{0ij} is the mean of the j^{th} subcluster in the i^{th} cluster, β_{1ij} is the treatment effect in the j^{th} subcluster in the i^{th} cluster, T_{ijk} is a treatment indicator coded $T_{ijk} = ½$ for treatment individuals and $T_{ijk} = -½$ for the control individuals, and the ε_{ijk} are independently normally distributed with mean 0 and variance σ_W^2. The level 2 (subcluster-level) model is

$$\beta_{0ij} = \gamma_{00i} + \eta_{0ij}, j = 1, ..., p, i = 1, ..., m$$

and

$$\beta_{1ij} = \gamma_{10i} + \eta_{1ij}, j = 1, ..., p, i = 1, ..., m,$$

where γ_{00i} is the mean of the i^{th} cluster, γ_{10i} is the treatment effect in the i^{th} cluster, the η_{0ij} are independently normally distributed with mean 0 and variance σ_{CBl}^2, and the η_{1ij} are independently normally distributed with mean 0 and variance $2\sigma_{T\times C}^2$. The level 3 (cluster-level) model is

$$\gamma_{00i} = \pi_{00} + \xi_{0i}, i = 1, ..., m$$

and

$$\gamma_{01i} = \pi_{10} + \xi_{1i}, i = 1, ..., m,$$

where π_{00} is the grand mean, π_{10} is the average treatment effect, the ξ_{0i} are independently normally distributed with mean 0 and variance σ_{SBl}^2, and the ξ_{1i} are independently normally distributed with mean 0 and variance $\sigma_{T\times S}^2$.

The treatment effect size δ is defined in terms of the unadjusted variances—that is,

$$\delta = \frac{\pi_{10}}{\sqrt{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}}\text{, where } \sigma_C^2 = \sigma_{CBl}^2 + \sigma_{T\times C}^2 \text{ and } \sigma_S^2 = \sigma_{SBl}^2 + \sigma_{T\times S}^2.$$

The ICCs are defined in terms of the unadjusted variances as

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}$$

and

$$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2},$$

and the heterogeneity parameters ω_S and ω_C are defined in terms of the unadjusted variances as

$$\omega_S = \frac{\sigma_{T\times S}^2}{\sigma_S^2}$$

and

$$\omega_C = \frac{\sigma_{T\times C}^2}{\sigma_C^2}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\pi_{10} = 0$.

Three-Level Randomized-Block Designs: Assigning Individuals Within Subclusters With Covariates

Suppose that there are m clusters, each with p subclusters of size $2n$, and that half of the individuals in each subcluster are assigned to each treatment. Now suppose that there are also $q \geq 0$ covariates at the cluster level, $r \geq 0$ covariates at the subcluster level, and $s \geq 0$ covariates at the individual level. Let Y_{ijk} be the k^{th} observation in the j^{th} subcluster of the i^{th} cluster. Thus, the level 1 (individual-level) model is

$$Y_{ijk} = \beta_{0ij} + \beta_{1ij}T_{ijk} + \beta_2 X_{2ijk} + \cdots + \beta_{q+1} X_{qijk} + \varepsilon_{ijk}\,,\ i = 1, \ldots, m; j = 1, \ldots p; k = 1, \ldots 2n$$

where β_{0ij} is the covariate-adjusted mean of the j^{th} subcluster in the i^{th} cluster, β_{1ij} is the covariate-adjusted treatment effect in the j^{th} subcluster in the i^{th} cluster, T_{ijk} is a treatment indicator coded $T_{ijk} = ½$ for treatment individuals and $T_{ijk} = -½$ for the control individuals, β_2, …, β_{q+1} are the effects of the individual-level covariates (which are fixed effects), X_{1ijk}, …, X_{qijk} are the values of the individual-level covariates (centered on subcluster means), and the ε_{ijk} are independently normally distributed with mean 0 and variance σ_{AW}^2. The level 2 (subcluster-level) model is

$$\beta_{0ij} = \gamma_{00i} + \gamma_{01} Z_{1ij} + \cdots + \gamma_{0r} Z_{rij} + \eta_{0ij}, j = 1, ..., p, i = 1, ..., m,$$

and

$$\beta_{1ij} = \gamma_{10i} + \gamma_{11} Z_{1ij} + \cdots + \gamma_{1r} Z_{rij} + \eta_{1ij}, j = 1, ..., p, i = 1, ..., m,$$

where γ_{00i} is the covariate-adjusted mean of the i^{th} cluster, γ_{10i} is the covariate-adjusted treatment effect in the i^{th} cluster, γ_{01}, …, γ_{0r} are the effects of the level 2 covariates (which are fixed effects) on the subcluster means, γ_{11}, …, γ_{1r} are the effects of the level 2 covariates on the subcluster-specific treatment effects, Z_{1ij}, … , Z_{rij} are the values of the subcluster-level

covariates (centered on cluster means), the η_{0ij} are independently normally distributed with mean 0 and variance σ_{ACBl}^2, and the η_{1ij} are independently normally distributed with mean 0 and variance σ_{ATxC}^2. The level 3 (cluster-level) model is

$$\gamma_{00i} = \pi_{00} + \pi_{01}W_{1i} + \ldots + \pi_{0s}W_{si} + \xi_{0i},\ i = 1, ..., m,$$

and

$$\gamma_{01i} = \pi_{10} + \pi_{11}W_{1i} + \ldots + \pi_{0s}W_{si} + \xi_{1i},\ i = 1, ..., m,$$

where π_{00} is the covariate-adjusted grand mean, π_{10} is the average covariate-adjusted treatment effect, $\pi_{01}, \ldots, \pi_{0s}$ are the effects of the level 3 covariates on the cluster means, $\pi_{11}, \ldots, \pi_{1s}$ are the effects of the level 3 covariates on the cluster-specific treatment effects, the ξ_{0i} are independently normally distributed with mean 0 and variance σ_{ASBl}^2, and the ξ_{1i} are independently normally distributed with mean 0 and variance σ_{ATxS}^2.

The treatment effect size δ is defined in terms of the unadjusted variances—that is,

$$\delta = \frac{\pi_{10}}{\sqrt{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}}$$

The ICCs are defined in terms of the unadjusted variances as

$$\rho_S = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2},$$

and

$$\rho_C = \frac{\sigma_C^2}{\sigma_S^2 + \sigma_C^2 + \sigma_W^2},$$

and the heterogeneity parameters ω_S and ω_C are defined in terms of the unadjusted variances as

$$\omega_S = \frac{\sigma_{T\times S}^2}{\sigma_S^2}$$

and

$$\omega_C = \frac{\sigma_{T\times C}^2}{\sigma_C^2}.$$

The covariate outcome correlations are defined in terms of the adjusted and unadjusted variances as

$$R_{T\times S}^2 = 1 - \frac{\sigma_{AT\times S}^2}{\sigma_{T\times S}^2},$$

$$R^2_{T\times C} = 1 - \frac{\sigma^2_{AT\times C}}{\sigma^2_{T\times C}},$$

and

$$R^2_{W} = 1 - \frac{\sigma^2_{AW}}{\sigma^2_{W}}.$$

The test for the treatment effect in this design is a test of the hypothesis that $\pi_{10} = 0$.

Appendix E: Glossary of Terms

Page #	Terms:
20	Cluster-level covariate: A covariate that is measured at the cluster (e.g., school or classroom) level. Please see also entry for covariate."
4	Clustered sampling: A technique in which a sample of naturally occurring groups called clusters (such as schools or residential blocks) are first selected, and then individuals are sampled within the selected clusters. Clustered sampling differs from stratified sampling in that individuals are selected from only some of the groups (clusters) in the population, whereas in stratified samples, individuals are intentionally selected from within *every* group (stratum).
13	Cohen's d: The standardized effect size computed by subtracting the mean of the control group from the mean of the treatment group, and dividing by a within-group standard deviation. Note that in multilevel designs different definitions of Cohen's d are possible (see Hedges, 2007). The current paper divides the mean difference by the total standard deviation to compute Cohen's d. Please also see the term "effect size."
5	Conditional inference model: The appropriate statistical model to use when conclusions from the data gathered are meant to apply only to the particular clusters and subclusters (eg. schools and classrooms) actually studied in the experiment. Appropriate only when the population of interest is fixed in both time and place. Please see also "Unconditional inference model."
12	Covariate: A variable that cannot be affected by the treatment and is expected to be correlated with the dependent variable. Ideally covariates are measured before the treatment is implemented. Covariates can be used to increase power in multilevel designs by decreasing residual variation between and within clusters.
7	Design effect (common usage): The ratio of the variance of an estimator under a particular sampling design to the variance of that estimator under a simple random sampling design.
14	Design effect (current paper): The gain in precision from sampling more than one unit per cluster or subcluster.
2	Effect size: A measure of the strength of the relationship between two variables. The current paper uses a version of Cohen's *d* as the effect size measure for computing statistical power.
N/A	Experiment: A research study where study subjects are exposed to conditions manipulated by the researcher, usually with the goal of determining the causal impact of some stimuli.

Page #	Terms:
N/A	False negative: To fail to identify a treatment effect when one exists or to erroneously fail to reject the null hypothesis—also known as Type II error.
N/A	False positive: To incorrectly detect a treatment effect when none exists or to erroneously reject the null hypothesis—also known as a Type I error.
34	Heterogeneity parameter (omega): One half of the variation in the treatment effects across clusters divided by the total variation across clusters. It is used in power computations for randomized-block designs.
9	Hierarchical design: An experimental design in which entire clusters (e.g., schools or classrooms) are assigned to treatments. Thus, every student in a given cluster (a school or classroom) receives the same treatment.
20	Individual-level covariate: A covariate defined at the individual level of a design.
1	Intraclass correlation coefficient (ICC): A parameter that measures the extent to which members of the same cluster or subcluster are more similar to one another than they are to members of other clusters or subclusters. In education research, for instance, an ICC may measure the extent to which the students in a particular classroom (or school) are more alike than students in another classroom (or school).
3	Nesting: Refers to the idea that certain units are contained within other units (for instance, schools are nested within school districts, classrooms are nested within schools, students are nested within classrooms).
18	Noncentral t-distribution: The sampling distribution of the t statistic when the null hypothesis is false. It has two parameters, degrees of freedom and the noncentrality parameter. It is used in statistical power calculations.
14	Noncentrality parameter: A quantity determining the distribution of a test statistic when the null hypothesis is false. It is the quantity most directly related to the statistical power of a test under a given alternative hypothesis.
2	Operational effect size: A modification to the usual effect size that can be used to compute power in multi-level designs using power tables intended for single level designs. The operational effect size is the usual effect size multiplied by a design effect that depends on features of the complex experimental design, such as the ICC.
13	Operational sample size: A modified sample size that can be used to compute power in multi-level designs using power tables intended for single level designs. The operational sample size is closely related to the number of clusters in the experiment.

Page #	Terms:
13	Optimal Design: A software program created by Raudenbush, Spybrook, Congdon, and Liu that is used for computing statistical power in group-randomized designs.
1	Power analysis: The calculation of the statistical power of a proposed research design for the purposes of ensuring that the design under consideration has a high enough chance of rejecting the a null hypothesis, when a given true effect of treatment is present.
13	Power table: A table that lists statistical power as a function of sample size and effect size.
10	Quasi-experiment: A research design that compares groups but does not involve randomization. Rather the treatment and comparison groups are often matched based on pretests or demographic factors, such as socioeconomic status.
10	Randomized-block designs): A class of experimental designs where units within the same cluster are randomly assigned to different treatments. For example, if students within the same school were randomly assigned to either of two treatments (e.g., an intervention and a control), the design is a two level randomized-block design.
11	Significance level: The probability of rejecting the null hypothesis when it is true—that is, the probability of a Type I error. The significance level 0.05 is often used in statistics.
N/A	Statistical inference: The process of deriving some conclusion about a population based on a sample.
1	Statistical power: The probability that the test of the null hypothesis of no average treatment effect will successfully reject the null hypothesis when a non-zero treatment effect exists.
4	Stratified sampling: A sampling method in which at least one individual from every one of an identified set of subgroups (e.g., schools) in a given population is intentionally included in the sample. This differs from clustered sampling, in which some subgroups (called clusters, in this case) have no individuals in the sample.
1	Type I error: Rejecting the null hypothesis when there is no treatment effect.
1	Type II error: Failing to reject the null hypothesis when a treatment effect is present.

Page #	Terms:
5	Unconditional inference model: A statistical model appropriate for use when the data in an experiment will be used to generalize to a population of interest that may not be fixed in either time or place. For example, this model is appropriate if the teachers and classrooms under study are considered a sample of potential teachers that could have been assigned to teach the students under study.
N/A	Unity: The number one.
7	Variance inflation factor: Please see the term “design effect (common usage).”

References

Bloom, H.S., Richburg-Hayes, L., and Black, A.R. (2007). Using Covariates to Improve Precision: Empirical Guidelines for Studies That Randomize Schools to Measure the Impacts of Educational Interventions. *Educational Evaluation and Policy Analysis*, *29*: 30-59.

Borenstein, M., Rothstein, H., and Cohen, J. (2001). *Power and Precision*. Teaneck, NJ: Biostat.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Academic Press.

Cooper H, and Hedges L. (1994). *The Handbook of Research Synthesis*. New York: Russel Sage Foundation.

Donner, A., and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Hedges, L.V., and Rhoads, C. (2009). *Statistical Power Analysis in Education Research.* U.S. Department of Education. Washington, DC: National Center for Special Education Research, Institute of Education Sciences.

Hedges, L.V., and Vevea, J.L. (1998). Fixed- and Random-Effects Models in Meta-Analysis. *Psychological Methods, 3*: 486-504.

Hedges, L.V., and Hedberg, E.C. (2007). Intraclass Correlations for Planning Group-Randomized Experiments in Education. *Educational Evaluation and Policy Analysis*, *29*: 60-87.

Johnson, N., and Kotz, S. (1971). *Distributions in Statistics: Continuous Univariate Distributions* (Vol. 1). New York: John Wiley.

Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley.

Konstantopoulos, S. (2008a). The Power of the Test for Treatment Effects in Three-Level Block Randomized Designs. *Journal of Research on Educational Effectiveness*, *1*: 265-288.

Konstantopoulos, S. (2008b). The Power of the Test for Treatment Effects in Three-Level Cluster Randomized Designs. *Journal of Research on Educational Effectiveness*, *1*: 66-88.

Murray, D.M. (1998). *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Nye, B., Hedges, L.V., and Konstantopoulos, S. (2000). The Effects of Small Classes on Achievement: The Results of the Tennessee Class Size Experiment. *American Educational Research Journal*, *37*: 123-151.

Raudenbush, S.W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Experiments. *Psychological Methods*, *2*: 173-185.

Raudenbush, S.W., and Liu, X. (2000). Statistical Power and Optimal Design for Multisite Randomized Trials. *Psychological Methods, 5*(3): 199-213.

Raudenbush, S., Martinez, A., and Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, *29*(1): 5-29.

Raudenbush, S., Spybrook, J., Congdon, R., and Liu, X. (2006). *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. William T. Grant Foundation.

Schochet, P. (2005). *Statistical Power for Random Assignment Evaluations of Education Programs*. U.S. Department of Education. Washington, DC: Institute of Education Sciences,

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Snijders, T. (2005). Power and Sample Size in Multilevel Modeling. In B.S. Everitt and D.C. Howell (Eds.), *Encyclopedia of Statistics in the Behavioral Sciences* (pp. 1570-1573). Chichester, England: Wiley.

Table 1: Power of the Test for Treatment Effects in the Hierarchical Design and a Function of Operational Sample Size N^T and Operational Effect Size Δ^T

	Effect size Δ^T																			
N^T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
3	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.08	0.08	0.09	0.09	0.10	0.10	0.11	0.11
4	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.09	0.10	0.10	0.11	0.13	0.14	0.15	0.16	0.17	0.19	0.20	0.22
5	0.05	0.05	0.06	0.06	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.16	0.18	0.20	0.22	0.24	0.27	0.29	0.32	0.34
6	0.05	0.05	0.06	0.07	0.08	0.09	0.10	0.12	0.14	0.16	0.18	0.21	0.23	0.26	0.29	0.33	0.36	0.39	0.43	0.46
7	0.05	0.06	0.06	0.07	0.08	0.10	0.12	0.14	0.16	0.19	0.22	0.25	0.29	0.32	0.36	0.40	0.44	0.49	0.53	0.57
8	0.05	0.06	0.07	0.08	0.09	0.11	0.13	0.16	0.19	0.22	0.26	0.30	0.34	0.38	0.43	0.48	0.52	0.57	0.61	0.66
9	0.05	0.06	0.07	0.08	0.10	0.12	0.15	0.18	0.22	0.25	0.30	0.34	0.39	0.44	0.49	0.54	0.59	0.64	0.69	0.73
10	0.05	0.06	0.07	0.09	0.11	0.13	0.16	0.20	0.24	0.29	0.33	0.39	0.44	0.49	0.55	0.60	0.65	0.70	0.75	0.79
11	0.05	0.06	0.07	0.09	0.12	0.14	0.18	0.22	0.27	0.32	0.37	0.43	0.49	0.54	0.60	0.66	0.71	0.76	0.80	0.84
12	0.05	0.06	0.08	0.10	0.12	0.16	0.20	0.24	0.29	0.35	0.41	0.47	0.53	0.59	0.65	0.71	0.76	0.80	0.84	0.88
13	0.05	0.06	0.08	0.10	0.13	0.17	0.21	0.26	0.32	0.38	0.44	0.51	0.57	0.63	0.69	0.75	0.80	0.84	0.88	0.91
14	0.05	0.06	0.08	0.11	0.14	0.18	0.23	0.28	0.34	0.41	0.47	0.54	0.61	0.67	0.73	0.78	0.83	0.87	0.90	0.93
15	0.05	0.06	0.08	0.11	0.15	0.19	0.24	0.30	0.37	0.43	0.50	0.58	0.64	0.71	0.77	0.82	0.86	0.90	0.92	0.95
16	0.05	0.07	0.09	0.12	0.15	0.20	0.26	0.32	0.39	0.46	0.54	0.61	0.68	0.74	0.80	0.84	0.88	0.92	0.94	0.96
17	0.05	0.07	0.09	0.12	0.16	0.21	0.27	0.34	0.41	0.49	0.56	0.64	0.71	0.77	0.82	0.87	0.91	0.93	0.96	0.97
18	0.05	0.07	0.09	0.13	0.17	0.22	0.29	0.36	0.43	0.51	0.59	0.67	0.74	0.80	0.85	0.89	0.92	0.95	0.97	0.98
19	0.05	0.07	0.09	0.13	0.18	0.23	0.30	0.38	0.46	0.54	0.62	0.69	0.76	0.82	0.87	0.91	0.94	0.96	0.97	0.98
20	0.06	0.07	0.10	0.14	0.19	0.25	0.32	0.40	0.48	0.56	0.64	0.72	0.78	0.84	0.89	0.92	0.95	0.97	0.98	0.99
21	0.06	0.07	0.10	0.14	0.19	0.26	0.33	0.41	0.50	0.58	0.67	0.74	0.81	0.86	0.90	0.94	0.96	0.97	0.98	0.99
22	0.06	0.07	0.10	0.15	0.20	0.27	0.35	0.43	0.52	0.61	0.69	0.76	0.83	0.88	0.92	0.95	0.97	0.98	0.99	0.99
23	0.06	0.07	0.11	0.15	0.21	0.28	0.36	0.45	0.54	0.63	0.71	0.78	0.84	0.89	0.93	0.96	0.97	0.98	0.99	1.00
24	0.06	0.08	0.11	0.16	0.22	0.29	0.37	0.47	0.56	0.65	0.73	0.80	0.86	0.91	0.94	0.96	0.98	0.99	0.99	1.00
25	0.06	0.08	0.11	0.16	0.22	0.30	0.39	0.48	0.58	0.67	0.75	0.82	0.88	0.92	0.95	0.97	0.98	0.99	1.00	1.00

26	0.06	0.08	0.11	0.17	0.23	0.31	0.40	0.50	0.60	0.69	0.77	0.84	0.89	0.93	0.96	0.97	0.99	0.99	1.00	1.00
27	0.06	0.08	0.12	0.17	0.24	0.32	0.42	0.52	0.61	0.70	0.78	0.85	0.90	0.94	0.96	0.98	0.99	0.99	1.00	1.00
28	0.06	0.08	0.12	0.17	0.25	0.33	0.43	0.53	0.63	0.72	0.80	0.86	0.91	0.95	0.97	0.98	0.99	1.00	1.00	1.00
29	0.06	0.08	0.12	0.18	0.25	0.34	0.44	0.55	0.65	0.74	0.81	0.88	0.92	0.95	0.97	0.99	0.99	1.00	1.00	1.00
30	0.06	0.08	0.12	0.18	0.26	0.35	0.46	0.56	0.66	0.75	0.83	0.89	0.93	0.96	0.98	0.99	0.99	1.00	1.00	1.00
31	0.06	0.08	0.13	0.19	0.27	0.37	0.47	0.58	0.68	0.77	0.84	0.90	0.94	0.96	0.98	0.99	1.00	1.00	1.00	1.00
32	0.06	0.09	0.13	0.19	0.28	0.38	0.48	0.59	0.69	0.78	0.85	0.91	0.94	0.97	0.98	0.99	1.00	1.00	1.00	1.00
33	0.06	0.09	0.13	0.20	0.29	0.39	0.50	0.61	0.71	0.79	0.86	0.92	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00
34	0.06	0.09	0.14	0.20	0.29	0.40	0.51	0.62	0.72	0.81	0.87	0.92	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00
35	0.06	0.09	0.14	0.21	0.30	0.41	0.52	0.63	0.73	0.82	0.88	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00
36	0.06	0.09	0.14	0.21	0.31	0.42	0.53	0.65	0.75	0.83	0.89	0.94	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00
37	0.06	0.09	0.14	0.22	0.32	0.43	0.54	0.66	0.76	0.84	0.90	0.94	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
38	0.06	0.09	0.15	0.22	0.32	0.44	0.56	0.67	0.77	0.85	0.91	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
39	0.06	0.09	0.15	0.23	0.33	0.45	0.57	0.68	0.78	0.86	0.92	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
40	0.06	0.09	0.15	0.23	0.34	0.46	0.58	0.69	0.79	0.87	0.92	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
41	0.06	0.10	0.16	0.24	0.35	0.47	0.59	0.70	0.80	0.88	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
42	0.06	0.10	0.16	0.24	0.35	0.48	0.60	0.72	0.81	0.89	0.94	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
43	0.06	0.10	0.16	0.25	0.36	0.48	0.61	0.73	0.82	0.89	0.94	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
44	0.06	0.10	0.16	0.25	0.37	0.49	0.62	0.74	0.83	0.90	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
45	0.06	0.10	0.17	0.26	0.37	0.50	0.63	0.75	0.84	0.91	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
46	0.06	0.10	0.17	0.26	0.38	0.51	0.64	0.76	0.85	0.91	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
47	0.06	0.10	0.17	0.27	0.39	0.52	0.65	0.77	0.86	0.92	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48	0.06	0.10	0.17	0.27	0.40	0.53	0.66	0.77	0.86	0.92	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
49	0.06	0.11	0.18	0.28	0.40	0.54	0.67	0.78	0.87	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	0.06	0.11	0.18	0.28	0.41	0.55	0.68	0.79	0.88	0.93	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
52	0.06	0.11	0.19	0.29	0.42	0.56	0.70	0.81	0.89	0.94	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
54	0.07	0.11	0.19	0.30	0.44	0.58	0.71	0.82	0.90	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
56	0.07	0.11	0.20	0.31	0.45	0.60	0.73	0.84	0.91	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

58	0.07	0.12	0.20	0.32	0.46	0.61	0.75	0.85	0.92	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
60	0.07	0.12	0.21	0.33	0.48	0.63	0.76	0.86	0.93	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
62	0.07	0.12	0.21	0.34	0.49	0.64	0.77	0.87	0.94	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.07	0.12	0.22	0.35	0.50	0.66	0.79	0.88	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
66	0.07	0.13	0.22	0.36	0.52	0.67	0.80	0.89	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
68	0.07	0.13	0.23	0.37	0.53	0.68	0.81	0.90	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
70	0.07	0.13	0.24	0.38	0.54	0.70	0.82	0.91	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
72	0.07	0.13	0.24	0.39	0.55	0.71	0.83	0.92	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
74	0.07	0.14	0.25	0.40	0.56	0.72	0.84	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
76	0.07	0.14	0.25	0.41	0.58	0.73	0.85	0.93	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
78	0.07	0.14	0.26	0.41	0.59	0.74	0.86	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
80	0.07	0.14	0.26	0.42	0.60	0.75	0.87	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
82	0.07	0.15	0.27	0.43	0.61	0.77	0.88	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
84	0.07	0.15	0.27	0.44	0.62	0.78	0.89	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
86	0.07	0.15	0.28	0.45	0.63	0.79	0.89	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
88	0.07	0.15	0.29	0.46	0.64	0.79	0.90	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
90	0.08	0.16	0.29	0.47	0.65	0.80	0.91	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Power of the Test for Treatment Effects in the Randomized-Block Design and a Function of Operational Sample Size N^T and Operational Effect Size Δ^T

	Effect size Δ^T																			
N^T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
2	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.09	0.09	0.10	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.17	0.18
3	0.05	0.06	0.06	0.07	0.08	0.10	0.12	0.13	0.16	0.18	0.20	0.23	0.26	0.29	0.32	0.35	0.38	0.41	0.44	0.47
4	0.05	0.06	0.07	0.09	0.11	0.14	0.17	0.21	0.25	0.29	0.33	0.38	0.43	0.48	0.53	0.58	0.63	0.67	0.72	0.75
5	0.05	0.06	0.08	0.11	0.14	0.18	0.23	0.28	0.34	0.40	0.47	0.53	0.59	0.65	0.71	0.76	0.81	0.85	0.88	0.91
6	0.05	0.07	0.09	0.13	0.17	0.22	0.29	0.36	0.43	0.51	0.58	0.66	0.72	0.78	0.83	0.88	0.91	0.94	0.96	0.97
7	0.06	0.07	0.10	0.15	0.20	0.27	0.34	0.43	0.52	0.60	0.68	0.75	0.82	0.87	0.91	0.94	0.96	0.97	0.98	0.99
8	0.06	0.08	0.11	0.17	0.23	0.31	0.40	0.50	0.59	0.68	0.76	0.83	0.88	0.92	0.95	0.97	0.98	0.99	1.00	1.00
9	0.06	0.08	0.13	0.19	0.26	0.35	0.46	0.56	0.66	0.75	0.82	0.88	0.93	0.96	0.97	0.99	0.99	1.00	1.00	1.00
10	0.06	0.09	0.14	0.21	0.29	0.40	0.51	0.62	0.72	0.80	0.87	0.92	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00
11	0.06	0.09	0.15	0.22	0.32	0.44	0.55	0.67	0.77	0.85	0.91	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
12	0.06	0.10	0.16	0.24	0.35	0.47	0.60	0.71	0.81	0.88	0.93	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
13	0.06	0.10	0.17	0.26	0.38	0.51	0.64	0.75	0.85	0.91	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	0.06	0.11	0.18	0.28	0.41	0.55	0.68	0.79	0.88	0.93	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	0.07	0.11	0.19	0.30	0.44	0.58	0.71	0.82	0.90	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	0.07	0.12	0.20	0.32	0.46	0.61	0.74	0.85	0.92	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	Effect size Δ^T																			
N^T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
17	0.07	0.12	0.21	0.34	0.49	0.64	0.77	0.87	0.94	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	0.07	0.13	0.22	0.36	0.52	0.67	0.80	0.89	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
19	0.07	0.13	0.24	0.38	0.54	0.70	0.82	0.91	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	0.07	0.14	0.25	0.40	0.56	0.72	0.84	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
21	0.07	0.14	0.26	0.42	0.59	0.74	0.86	0.94	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	0.07	0.15	0.27	0.43	0.61	0.77	0.88	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
23	0.07	0.15	0.28	0.45	0.63	0.79	0.89	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
24	0.08	0.16	0.29	0.47	0.65	0.80	0.91	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25	0.08	0.16	0.30	0.48	0.67	0.82	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
26	0.08	0.17	0.31	0.50	0.69	0.84	0.93	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
27	0.08	0.17	0.32	0.52	0.71	0.85	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
28	0.08	0.18	0.33	0.53	0.72	0.86	0.95	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
29	0.08	0.18	0.34	0.55	0.74	0.88	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30	0.08	0.19	0.36	0.56	0.75	0.89	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
31	0.08	0.19	0.37	0.58	0.77	0.90	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
32	0.09	0.20	0.38	0.59	0.78	0.91	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
33	0.09	0.20	0.39	0.61	0.80	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	Effect size Δ^T																			
N^T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
34	0.09	0.20	0.40	0.62	0.81	0.92	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
35	0.09	0.21	0.41	0.63	0.82	0.93	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
36	0.09	0.21	0.42	0.65	0.83	0.94	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
37	0.09	0.22	0.43	0.66	0.84	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
38	0.09	0.22	0.44	0.67	0.85	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
39	0.09	0.23	0.45	0.68	0.86	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40	0.09	0.23	0.46	0.69	0.87	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
41	0.10	0.24	0.47	0.71	0.88	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
42	0.10	0.24	0.48	0.72	0.89	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
43	0.10	0.25	0.48	0.73	0.89	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
44	0.10	0.25	0.49	0.74	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
45	0.10	0.26	0.50	0.75	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
46	0.10	0.26	0.51	0.76	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
47	0.10	0.27	0.52	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48	0.10	0.27	0.53	0.77	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
49	0.11	0.28	0.54	0.78	0.93	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	0.11	0.28	0.55	0.79	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	Effect size Δ^T																			
N^T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
51	0.11	0.29	0.56	0.80	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
52	0.11	0.29	0.56	0.81	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
53	0.11	0.30	0.57	0.82	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
54	0.11	0.30	0.58	0.82	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
55	0.11	0.31	0.59	0.83	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
56	0.11	0.31	0.60	0.84	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
57	0.12	0.32	0.60	0.84	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
58	0.12	0.32	0.61	0.85	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
59	0.12	0.33	0.62	0.86	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
60	0.12	0.33	0.63	0.86	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
61	0.12	0.34	0.64	0.87	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
62	0.12	0.34	0.64	0.87	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
63	0.12	0.35	0.65	0.88	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
64	0.12	0.35	0.66	0.88	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
65	0.12	0.36	0.66	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
66	0.13	0.36	0.67	0.89	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
67	0.13	0.36	0.68	0.90	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

	Effect size Δ^T																			
N^T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
68	0.13	0.37	0.68	0.90	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
69	0.13	0.37	0.69	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
70	0.13	0.38	0.70	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00