Optimal Design Plus Empirical Evidence:
Documentation for the "Optimal Design" Software

Jessaca Spybrook[a]

Howard Bloom[b] Richard Congdon[c] Carolyn Hill[d] Andres Martinez[e] Stephen Raudenbush[f]

APPLIES TO: Optimal Design Plus Version 3.0

LAST REVISED ON: October 16, 2011

This work was funded by the William T. Grant Foundation. We would like to thank the foundation for its continued support for the Optimal Design Plus Software and documentation. We also want to thank Xiaofeng Liu for his help calculating the power for the various designs included in the software and documentation.

[a] Western Michigan University
[b] MDRC
[c] NORC
[d] Georgetown University
[e] University of Michigan
[f] University of Chicago

Table of Contents

Section I: Introduction

This manual describes how to conduct a power analysis for individual and group randomized trials. The manual includes an overview of each design, the appropriate statistical model, and, for each, the calculation of statistical power and minimum detectable effect size. The manual also explains how to use the Optimal Design Software Version 3.0 for planning adequately powered experiments. The manual is divided into 6 sections. Section 1 provides a brief introduction to power analysis, describes the various designs available in the software, and describes the setup of the software. We recommend that users read Section 1 first to understand the main features of the program. Sections 2 through 6 are stand-alone chapters that are specific to particular research designs and outcome types. Appendix A provides resources for the Empirical Based MDES described in Section IV. Appendix B and C are stand-alone pieces that describe power for meta-analysis and optimal sample allocation for two-level cluster randomized trials.

1.0 Statistical power

Power is the probability of rejecting the null hypothesis when a specific alternative hypothesis is true. In a study comparing two groups, power is the chance of rejecting the null hypothesis that the two groups share a common population mean and therefore claiming that there *is* a difference between the population means of the two groups, when in fact there is a difference of a given magnitude. It is thus the chance of making the correct decision, that the two groups are different from each other. Power is linked to discussions of hypothesis testing and significance levels, so it is important to have a clear definition of each of these terms before proceeding. Note that in a perfectly implemented randomized experiment with correctly analyzed data, power is the probability of discovering a causal effect of treatment when such an effect truly exists.

In hypothesis testing, there are two hypotheses, a null hypothesis and an alternative hypothesis. In a two-treatment design, the most common null hypothesis states that there is no difference between the population means of the treatment and control groups on the outcome of interest. The alternative hypothesis states that there is a difference between groups. The difference may be expressed as a positive treatment effect, a negative treatment effect, or simply that the treatment mean is not equal to the control mean. After the hypotheses are clearly stated and the data have been collected and analyzed, the researcher must decide if there is sufficient evidence to reject the null hypothesis.

The significance level, often denoted α, is the probability of rejecting the null hypothesis when it is true. This is known as a Type I error rate. A Type I error occurs when the researcher finds a significant difference between two groups that do not, in fact, differ. Suppose, however, that the null hypothesis is indeed false. A Type II error arises when we mistakenly retain the null hypothesis. The probability of retaining a false null hypothesis, often denoted β, is therefore the Type II error rate. In this case, the researcher overlooks a significant difference. The two types of errors are illustrated in Table 1.1.

Table 1.1

Possible errors in hypothesis testing

	Do Not Reject the Null Hypothesis	**Reject the Null Hypothesis**
Null Hypothesis is True	No Error (Probability = 1-α)	Type I Error (Probability = α)
Null Hypothesis is False	Type II Error (Probability = β)	No Error (Probability = 1-β)

If the null hypothesis is true (first row of Table 1.1), the correct decision is to retain the null and the probability of this correct decision = Probability (Retain $H_0 \mid H_0$ is true) = 1-α. With $\alpha = 0.05$, for example, the probability is 0.95 that we will make the correct decision of retaining H_0 when it is true. The incorrect decision in this case is the Type I error – rejecting the true H_0. When H_0 is true, this error will occur with probability $\alpha = 0.05$.

On the other hand, if the null hypothesis is false (second row of Table 1.1) the correct decision is to reject it. If the probability of making this correct decision is defined as power = Probability (Reject $H_0 \mid H_0$ is false)=$1 - \beta$. The incorrect decision, known as the Type II error occurs with probability β, that is Prob(Type II error$\mid H_0$ false)= β.

Looking at the results of a study retrospectively, we know that a researcher who has retained H_0 (column 1 of Table 1.1) has either made a correct decision or committed a Type II error. In contrast, a researcher who has rejected H_0 (column 2) has either made a correct decision or committed a Type I error. Note that it is logically impossible for a researcher who has rejected H_0 to have made a Type II error. To criticize such a researcher for designing a study with low power in this case would be a vacuous criticism, because a lack of power cannot account for a decision to reject H_0. However, a researcher who retains the null hypothesis may have committed a Type II error and is therefore potentially vulnerable to the criticism that the study lacked power. Indeed, low power studies in which H_0 is retained are virtually impossible to interpret. One cannot claim a new treatment to be ineffective in a study having low power

because, by definition, such a low power study would have little chance of detecting a true difference between two populations represented in the study.

Although Type I and Type II errors are mutually exclusive, the choice of α can affect power. Suppose a researcher, worried about committing a Type I error, sets a lower α, say $\alpha = 0.001$. If the null hypothesis is true, this researcher will indeed be protected against a Type I error. However, suppose H_0 is false. Setting α very low will reduce power, equivalent to increasing β, the probability of a Type II error. While keeping in mind that the choice of α affects power, we will for simplicity assume $\alpha = 0.05$ in the remainder of this discussion in order to focus on sample size as a key determinant of power.

Of course, neither type of error is desirable and we would prefer to make the correct decision. As a result, we want the probability of correctly detecting a difference, that is, the power, to be large. For example, if the power is 0.80, we will correctly identify a difference between the groups with probability 0.80. Power greater than or equal to 0.80 is often recognized by the research community to be sufficient, though some researchers seek 0.90 as a minimum.

The ability to correctly detect a difference of a given magnitude in the mean outcome for the two groups is characterized by the power of the study. If a study is underpowered, a practically significant true difference might go undetected. The importance of designing a study with adequate power cannot be overstated, especially from the cost perspective. Imagine a multi-million dollar intervention study that fails to detect an effect simply because the study did not have sufficient power. In other words, the intervention may or may not produce practically significant effects, but the researchers are not able to make this determination due to inadequate power. One might argue that the money invested in the trial was not well spent since at the end of the study, it is still unclear whether or not the intervention was effective.

1.1 Approaches for conducting a power analysis

In the recent literature on statistical power, two approaches for conducting power analyses have emerged. The first approach, which we call the "power determination approach," begins with an assumption about the effect size the intervention produces, and the aim is to compute the power they will have to detect that effect with a given sample size. For example, suppose that a team of researchers is planning a study to detect the effect of a school-level intervention aimed at improving math achievement for third graders. They plan to randomize schools to receive either the treatment or continue with current protocol. Pilot studies and

available theory suggest that a practically significant effect would entail a standardized effect size of 0.20; that is, a mean difference equivalent to 0.20 in units of the population standard deviation of the outcome. Thus the researchers want to plan the study to be able to detect an effect of at least 0.20 standard deviation units. In this case, the effect size is already determined, and the researchers are interested in calculating the sample size necessary to achieve power of 0.80. Of course, this process can be repeated for a range of effect sizes.

The second approach, which we call the "effect size approach," begins with a desired level of power and the aim is to compute the minimum effect size that can be detected at that level of power for any given sample size. This approach can, of course, be replicated at any given level of power. Bloom (1995) defines the MDES as the smallest true effect that can be detected for a specified level of power and significance level for any given sample size. For example, suppose that another team of researchers is studying a whole school reform model. They plan to randomize schools to either the new reform model or current conditions. Because of financial considerations, the team can only recruit 50 schools and 100 students within each school. The sample size is set, thus the researchers are trying to determine the smallest effect size they can detect with the pre-specified sample size.

The power determination approach and the effect size approach represent two different ways to conduct a power analysis. However, both approaches yield the same conclusions. That is, a power analysis could be conducted using either approach and the ultimately the same conclusions would be reached. Also, both approaches also require assumptions about the variation in the outcome. The Optimal Design software allows the researcher to use either approach for conducting the power analysis.

2.0 Design options

Identifying the appropriate research design for a study is critical because a power analysis is specific to a particular design. That is, the required parameters differ depending on whether, for example, individuals are the unit of randomization or clusters are the unit of randomization, or blocking is or is not present. This section summarizes the various design options present in Optimal Design Version 3.0. The models and notation correspond to the HLM notation (Raudenbush and Bryk, 2002). Specific details about how to calculate the power for the various designs is found in sections 2, 3, 5, and 6.

Table 2.1 includes all of the design options when the primary outcome is measured at the individual level. The designs are divided into two groups, those that randomly assign individuals, hereafter referred to as person randomized trials, and those than randomly assign clusters, or intact groups of individuals, hereafter referred to as cluster randomized trials (CRT). The first row of the table identifies the number of levels in the study. For example, a single level trial simply has one level, whereas a multi-site trial can be conceived as a two level trial, with individual in sites or blocks. We can look at rows 2 through 4 together to understand the relationship between the level of randomization, the number of levels, and the presence of blocking. For the single level trial and the simple nested designs that do not include blocking, we can see that the level of randomization is the same as the top level in the study. For example, in a three level cluster randomized trial (3-level CRT), there are three levels and the top level, or level three, is the unit of randomization. In the blocked designs, the level of randomization is immediately below the blocks, with the blocks being the top level. For example, in a multi-site cluster randomized trial (MSCRT), there are three levels, possibly students in classrooms in schools and schools are blocks. Randomization occurs within the blocks, hence at level 2, or one level below the blocks. This is true for all the designs that include blocking in Table 2.1.

The next row indicates whether or not there is the option to include a covariate in the analysis in the software. In the cases where a covariate is available, the covariate is always at the level of randomization, with the exception of the Empirically Based MDES, which allows covariates to be included at all levels. Including a covariate is a common way to increase the precision of the study and thus reduce the required sample size, which can often help reduce the cost of the study. The use of a covariate requires that the following assumptions are met: 1) the

covariate has a strong linear association with the outcome, and 2) the association is similar within each treatment condition.

The row labeled "outcome" identifies the outcome type that the Optimal Design accepts. For the single level trials, continuous outcomes are the only available option. Power analysis for binary outcomes is available for three of the CRT's. For three of the designs we also have empirical estimates available within the program. The empirically based options are described in Section IV. The final row provides an example of the nested structure of the data for each design.

Table 2.1

Design Options for Individual Level Outcome Measures

	Person Randomized Trials			Group Randomized Trials				
	Single-level Trial	Multi-site (or blocked) Trial	Repeated Measures Trial	Two-Level Cluster Randomized Trial	Three-level Cluster Randomized Trial	Three-level Multi-site Cluster Randomized Trial[a]	Four-Level Multi-site Cluster Randomized Trial	Cluster Randomized Trial with Repeated Measures
Number of Levels	1	2	2	2	3	3	4	3
Level of Randomization	1	1	2	2	3	2	3	3
Blocking?	No	Yes	No	No	No	Yes	Yes	No
Covariate?	Yes	Yes	No	Yes[a]	Yes[a]	Yes[a]	Yes	No
Outcome type	Continuous	Continuous	Continuous	Continuous Binary	Continuous Binary	Continuous Binary	Continuous	Continuous
Empirical Estimates Available	No	No	No	Yes[b]	No	Yes[b]	Yes[b]	No
Example	Students	Students, Schools	Repeated measures for students	Students, Schools	Students, Classrooms, Schools	Students, Classrooms, Schools (blocks)	Students, Classroom, Schools, Districts (blocks)	Repeated measures for students, schools

[a] Option available for the continuous case only. [b] Covariates also allowed at lower levels.

The second set of design options available in the software includes designs in which the primary interest is in a group-level measure instead of an individual level measure. For example, a measure of classroom quality might be the primary outcome. We shall assume, however, that the group-level outcome is measured imperfectly, that is, with reliability less than 1.0. In this case measurement error variance "adds a level" to the analysis (see Raudenbush and Bryk (2000), Chapter 11. Table 2.2 presents these options. These designs look similar to those in Table 2.1 as far as the first four rows of the table. The main difference is that the outcome of interest is measured at the group level rather than the individual. This is evident by the absence of the individual in the examples of nesting in row 4 of Table 2.2.

Table 2.2

Design Options for Group Level Outcome Measures

	2-level cluster randomized trial	3-level cluster randomized trial	Multi-site cluster randomized trial
Number of Levels	2	3	3
Level of Randomization	2	3	2
Blocking?	No	No	Yes
Covariate?	No	Yes	No
Outcome type	Continuous	Continuous	Continuous
Example	Classrooms	Classrooms Schools	Classrooms Schools

Tables 2.1 and 2.2 identify the designs available in the OD software. One major design difference that emerges across the tables is whether or not a trial includes blocking. Although we leave the specific details of each design to Sections 2 through 5, we discuss the rationale for blocking since it applies across all the blocked designs identified in Tables 2.1 and 2.2.

2.1 Blocking

Blocking is a commonly used in experimental design to improve the face validity and/or to improve the precision and power of the experimental study. For person randomized trials, the basic idea of pre-randomization blocking is to find sites or blocks where individuals within the sites are very similar with respect to the outcome variable. One then randomly assigns persons to treatments within each block. Variation between blocks does not affect the standard error of the treatment effect estimate; if such variation is large, blocking will increase statistical power. The same idea extends to cluster randomized trials; then the aim is to find blocks where clusters are similar with respect to the outcome variable. This reduces the heterogeneity within blocks, increasing the precision of the treatment effect estimate, hence increasing the power of the test for the main effect of treatment. Researchers often regard the blocks as "sites," so that a cluster randomized trial with blocking is often defined as a "multi-site cluster randomized trial," and we shall use that language as well.

To illustrate in the case of a cluster randomized trial, imagine that researchers develop a new reading program for elementary school students. We know that the percent of students with free/reduced lunch is related to school mean reading achievement. We might therefore assign the school to "blocks" that are similar percent with free and reduced lunch. Within each block, we randomize schools to receive the new reading program or the regular program. This reduces the variance in the estimate of the treatment effect because by dividing schools into blocks we are able to remove the between-block variance from the error variance. If the between-block component is large, removing it greatly increases the precision of the estimate. Another example arises because schools are naturally grouped within school districts. The districts are then blocks or sites, and the randomization occurs within districts.

We define designs that block before randomizing as multi-site randomized trials. In essence, they are single level trials or cluster randomized trials that are being replicated within each site. Replication across sites allows us to estimate an effect size for each site. Thus we are able to estimate the variability of the treatment effect across sites.

In many cases, the sites will be regarded as randomly sampled from a larger universe or "population" of possible sites. The larger universe is the target of generalization. For example, if schools are sampled and then classrooms are assigned at random to treatments within schools,

the target of any generalizations will often be the larger universe of schools from which schools in the study are regarded as a representative sample.

In other cases, the sites will be regarded as fixed. Consider a program designed to teach students about the dangers of drugs. The outcome for the study is students' attitude towards drugs, which is measured by a questionnaire. The researchers hypothesize that the school setting - suburban, urban, or rural - affects students' attitude towards drugs. Thus they want to block on the setting. In this case, suburban, urban, and rural are not regarded as sampled from a population of settings, but rather as fixed blocks or sites. Whether we view sites as fixed or random affects the data analysis and planning for adequate power to detect the treatment effect.

3.0 Layout of the Optimal Design Software

This chapter describes the setup of the software. Section 3.1 describes how to navigate through the mainscreen of OD. Section 3.2 describes the layout of each module. A concluding section highlights the underlying assumptions in the software.

3.1 Navigating the main screen

The blank screen in the Optimal Design is displayed in Figure 3.1.

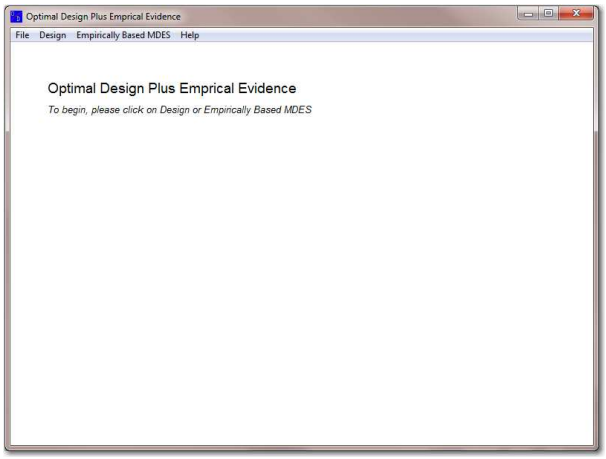


Figure 3.1. Initial blank screen.

Clicking on the file option reveals the preferences and the exit options. The preferences allow the user to select black and white or color for the graphs on the screen and any saved graphs. Also, the user can select to integer or continuous values on the horizontal axis[1]. Figure 3.2 displays the preferences screen.

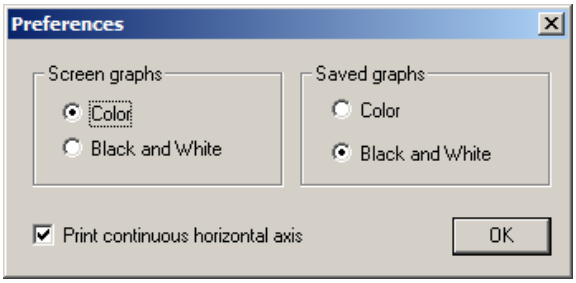


Figure 3.2. Preferences.

The help option provides the user with resources and contact information for the Optimal Design authors.

[1] Clicking along the trajectory is not an exact method for obtaining the power for a study. It gives a very close estimate. Selecting continuous values may help the user find a more exact value. R code is available upon request for users interested in the exact power.

There are two primary menu options: Design and Empirically Based MDES. The Design option displays graphical output for all of the designs described in Tables 2.1 and 2.2. The user is required to enter her own design parameters to complete the power analysis. The Empirically Based MDES displays tabular output. This option allows the user to access empirically based values for the design parameters for continuous outcomes. Currently this option is available for the 2-level CRT, the MSCRT with Treatment at Level 2, and the MSCRT with Treatment at Level 3. We discuss the Empirically Based MDES in Section 4. The remainder of this section describes the options under the design tab.

Clicking on the design tab brings up four options:

Design

Person randomized trial

Cluster randomized trial with person-level outcomes

Cluster randomized trial with cluster-level outcomes

Meta-analysis

We focus on the first three choices, however, the power for a meta-analysis is discussed in Appendix B. The first three choices, person randomized trials, cluster randomized trials with person-level outcomes, and cluster randomized trials with cluster-level outcomes, correspond to the main design options defined in Chapter 2. Within each type, there are various design choices as identified in Tables 2.1 and 2.2. The specific details for each design option are included in Sections 2, 3, 5, and 6. However, each module functions similarly and section 3.2 describes the general layout of each module.

3.2 General layout

The OD is setup to encourage the user to have already defined the design prior to running power calculations. That is, the user must navigate through a series of design prompts prior to reaching the screen which enables him to conduct a power analysis. The first thing the user must determine is whether the trial is a person randomized trials, a cluster randomized trial with individual outcomes, or a cluster randomized trial with group-level outcomes. We discuss each option separately.

Person Randomized Trials

Placing the mouse over the heading person randomized trials reveals three options:

Person Randomized Trial

Single level trial

Multi-site (or blocked) trials

Repeated measures

The design choices correspond to those in Table 2.1 and the user must select the appropriate design at this stage. After selecting a design, the main menu for the design will appear. The menu for each design varies slightly depending on the design and is described in detail in the individual design chapters.

Cluster Randomized Trials with person-level outcomes

Placing the mouse over the heading cluster randomized trials with person-level outcomes reveals two options:

Cluster Randomized Trials with person-level outcomes

Cluster randomized trials

Multi-site (or blocked) cluster randomized trials

After clicking on either a cluster randomized trial or blocked trial, the user is asked to specify the level of treatment. After selecting the level of treatment, the main menu for the design appears.

Cluster Randomized Trials with group-level outcomes

Placing the mouse over the heading cluster randomized trials with group-level outcomes reveals two options:

Cluster Randomized Trials with group-level outcomes

Cluster randomized trials

Multi-site (or blocked) cluster randomized trials

Similar to the CRT for person-level outcomes, the user is prompted to decide either a cluster randomized trial or a blocked trial. Selecting on the cluster randomized trial forces the user to select either treatment at level 2 or 3 in order to enter the main menu for a design. There is only one option for the blocked trial, treatment at level 2, and once selected, the user enters the main menu.

3.3 Power or MDES on y-axis

After navigating through the prompts to the appropriate design, the user must select a display option. The two primary choices are either Power on the y-axis or MDES on the y-axis. For either of these options, the main menu for each design is very similar. Figure 3.3 displays the

main menu for a person randomized trial → Single level trial → Power on y-axis → Power vs. total number of people.

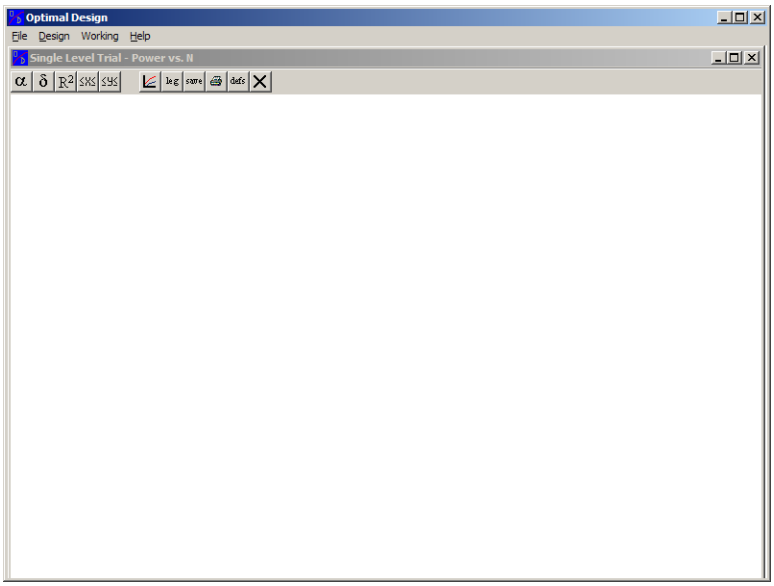


Figure 3.3. Main menu for a person randomized trial.

The buttons that appear at the top of the screen vary for each design depending on the parameters that are required for a power analysis. However, the general layout is the same. The title appears at the top, in this case, Power vs. total number of people (*N*). This indicates that the power will be on the y-axis and the total number of people will vary along the x-axis. Below is an explanation of the buttons that appear below the title.

α	is the significance level, or Type I error rate. By default, it is set to 0.05. It can be changed by clicking on it and changing the value.
δ, R^2	are the design parameters required for conducting a power analysis. For other designs, other parameters may be required. To set these parameters, the user simply clicks on the button and sets the value. The number of options for each parameter varies from 1 to 3.
$\leq X \leq$	controls the minimum and maximum values on the x-axis. The minimum and maximum values can be changed by clicking this button.
$\leq Y \leq$	controls the minimum and maximum values on the y-axis. By default, the y-axis is set from 0.0 to 1.0 but can be changed by clicking on the button Graph symbol plots the default settings.

Leg	allows the user to change the title of the x-axis legend, y-axis legend, and to add a title to the graph.
Save	allows the user to save the graph. Graphs are saved as .emf files and can be inserted into a word document using the insert picture command.
Print symbol	prints the graph on the screen.
Defs	plots the default settings.
X	closes the graph and returns the user to the original screen.

Clicking on any button automatically yields a power curve. Figure 3.4 is the default power curve for the single level trial.

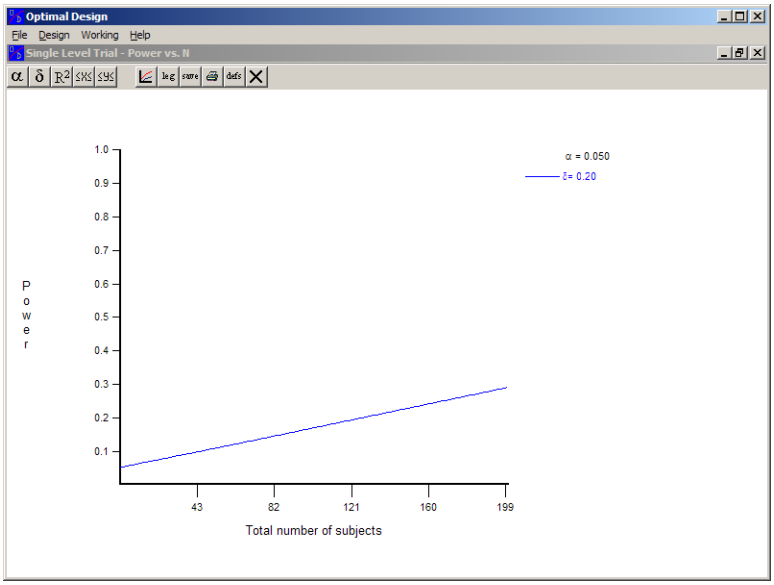


Figure 3.4. Default settings for single level trial.

The key appears in the upper right corner of the screen and lets the user know the specified parameters. The parameters are changed by clicking on the buttons. Clicking along the trajectory also allows the reader to determine the power for a specific sample size.

3.4 Assumptions

There are several assumptions underlying OD. First, for the traditional graphical output, we assume that all designs are balanced. For example, in a single level trial with 60 people, we assume that 30 people are in the treatment group and 30 are in the control group. In some cases, the design is purposely imbalanced or differences in cluster sizes are unavoidable. We recommend using the harmonic mean for these cases. The Empirically Based MDES allows for

unbalanced designs. In this case, the harmonic mean is calculated within the program for further power calculations.

A second assumption is that there are two conditions. In all cases, the power is calculated for the difference between two groups, treatment and control. For multiple groups, we recommend using the software to determine the power for pairwise comparisons.

A third assumption is that the parameters entered into the software are reasonable. The OD accepts all parameter values and does not test whether or not a parameter value is realistic. Pilot data and literature reviews are the most appropriate methods for obtaining reasonable parameters to use for a power analysis. The default values are simply default values and do not apply to any particular study. The data available in the Empirically Based MDES is not applicable for all studies and should be carefully considered before using it for power calculations.

Section II: Optimal Design for person randomized trials

Optimal Design for person randomized trials includes trials where individuals are randomly assigned to the treatment or control condition. There are three types of designs in this category. Briefly, single level trials are trials with no blocking or clustering. That is, individuals are randomly assigned to either the treatment or the control group. Multi-site (or blocked) trials are studies where individuals are randomly assigned to the treatment or control within blocks. That is, the randomization process is repeated across blocks or sites. The blocks may either be intact entities such as classrooms or they may be matched pairs, where individuals are put into pairs (or blocks) because they are similar with respect to a variable that is related to the outcome. The third option, repeated measures, are studies in which individuals are randomly assigned to the treatment or control and then the individuals are measured repeatedly over time. We describe the conceptual details and provide a "how to" guide for each design in the following 3 chapters.

4.0 Single level trials

Single level trials rely on the assignment of individuals to a treatment condition. In a single level design, we simply randomize individuals to a treatment condition or control condition. The use of random assignment assures that the treatment groups are comparable. First, we examine the statistical models to see what effects the power to detect the treatment effect.

4.1 The model

We can represent data from a single level trial with a simple one level model. The model can be expressed as:

$$Y_i = \beta_0 + \beta_1 W_i + r_i \qquad r_i \sim N(0, \sigma^2) \qquad [4.1]$$

for

$i=1,\ldots,N$ persons in the study

where

Y_i is the response for person i

β_0 is the mean response

β_1 is the treatment effect

W_i is the treatment indicator with ½ for treatment and -½ for control

r_i is the random error associated with each person i

σ^2 is the between persons variation.

4.2 Testing the treatment effect

We are primarily interested in the main effect of treatment, β_1, or in a balanced design, the simple difference between the treatment and control averages. It is estimated by:

$$\hat{\beta}_1 = \bar{Y}_E - \bar{Y}_C \qquad [4.2]$$

where

$\bar{Y}_E$ is the mean for the experimental group

$\bar{Y}_C$ is the mean for the control group.

Assuming $N/2$ persons per cluster, the variance of the estimated treatment effect is:

$$\text{var}(\hat{\beta}_1) = \frac{4\sigma^2}{N} \qquad [4.3]$$

Note that the variance of the treatment effect is a function of the total sample size, *N*, and the between-persons variance, σ^2.

We can use the results of a one way analysis of variance with a fixed effect for the treatment. The test statistic is an *F* statistic, which compares treatment variance to error variance. The *F* statistic is defined as:

$$Fstatistic = \frac{(MS_{treatment})}{(MS_{error})}. \qquad [4.4]$$

As *N* increases without bound, the *F* statistic converges to the ratio of expected mean squares, defined as:

$$\frac{E(MS_{treatment})}{E(MS_{error})} = \frac{\sigma^2 + N\beta_1^2/4}{\sigma^2} = 1 + \frac{N\beta_1^2}{4\sigma^2} \qquad [4.5]$$

and can be rewritten as:

$$\frac{E(MS_{treatment})}{E(MS_{error})} = 1 + \lambda \text{ where } \lambda = \frac{N\beta_1^2}{4\sigma^2} = \frac{\beta_1^2}{Var(\hat{\beta}_1)} \qquad [4.6]$$

If the null hypothesis is true, the *F* statistic follows a central *F* distribution with 1 degree of freedom for the numerator and *N*-2 degrees of freedom for the denominator. Under the central *F* distribution, we would expect the *F* statistic to be approximately 1. In other words, there is no variation between treatments so $\beta_1 \approx 0$ and the term with $N\beta_1^2$ in the numerator of the expected mean square ratio is null. We see that if $\lambda = 0$ the ratio of expected mean squares converges in large samples to $\frac{E(MS_{treatment})}{E(MS_{cluster})} = \frac{\sigma^2}{\sigma^2} = 1 + \lambda = 1.$

If the null hypothesis is false so that there is a treatment difference, that is $\beta_1 \neq 0$, the *F* statistic follows a non-central *F* distribution with 1 degree of freedom for the numerator and *N*-2 degrees of freedom for the denominator and non-centrality parameter. Then the ratio of expected mean squares becomes the non-central *F* distribution, characterized by a non-centrality parameter, λ, defined in Equation 4.6. Note that λ can also be expressed as the ratio of the squared treatment effect to the variance of the estimate of the treatment effect.

The non-centrality parameter is strongly related to the power of the test. As λ increases, the power increases. Looking at Equation 4.6, we can see that the non-centrality parameter is a function of *N*, σ^2, and β_1. As σ^2, the variation between-persons, decreases, the non-centrality parameter will increase. As the desired effect size β_1 increases, the non-centrality parameter increases. However, the problem with these two parameters is that they typically are not under the control of the researcher. The effect size and between-person variability are usually a function of the phenomenon under consideration. As a result, the most effective way for the researcher to increase the power of the test to detect the treatment effect of a given magnitude is to increase the total sample size, *N*. As *N* increases, the non-centrality parameter increases as well.

4.3 Standardized notation

Thus far we have focused on the unstandardized notation. However, it will be easier to think in terms of standardized units. A standardized effect size, δ, is the difference in the population means of the two groups divided by the standard deviation of the outcome. The standardized effect in a single level trial can be expressed as:

$$\delta = \frac{\beta_1}{\sqrt{\sigma^2}} \qquad [4.7]$$

where

$\beta_1 = \mu_E - \mu_C$

μ_E is the population mean for the experimental group

μ_C is the population mean for the control group.

In the standardized model, we set $\sigma^2 = 1$. Dividing the numerator and denominator of the non-centrality parameter by σ^2, we see that we can represent the non-centrality parameter in standardized notation simply as:

$$\lambda = \frac{N\beta_1^2 / \sigma^2}{4\sigma^2 / \sigma^2} = \frac{N\delta^2}{4}. \qquad [4.8]$$

This allows us to calculate the power as a function of only two parameters, the total sample size and the standardized effect size.

4.4 The model with a covariate

Choice of an effective pre-treatment predictor known as a "covariate" will reduce the between-person variation, hence increasing the precision of the estimate of the treatment effect. The correlation between the covariate and the outcome is denoted, ρ_{xy}. The proportion of variance explained by the covariate is denoted ρ_{xy}^2.

Equation 4.9 is the model with a covariate.

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + r_i \qquad r_i \sim N(0, \sigma_{|x}^2) \qquad [4.9]$$

where

Y_i is the response for person i

β_0 is the mean response

β_1 is the treatment effect

W_i is the treatment indicator with ½ for treatment and -½ for control

β_2 is the regression coefficient for the covariate

X_i is the value of the covariate, centered around its grand mean

r_i is the random error associated with each person i conditional on the covariate

$\sigma_{|x}^2$ is the between person variation conditional on the covariate, and can therefore be regarded as the "conditional variance," where $\sigma_{|x}^2 = (1 - \rho_{xy}^2)\sigma^2$.

Note that the model now looks like the familiar analysis of covariance model. The variance is conditional on the covariate. The smaller the conditional variance relative to the unconditional variance, the greater the increase in the precision of the treatment effect.

4.6 Testing the treatment effect (including a covariate)

The estimate of the treatment effect is:

$$\hat{\beta}_1 = \bar{Y}_E - \bar{Y}_C - \hat{\beta}_2(\bar{X}_E - \bar{X}_C). \qquad [4.10]$$

The estimate of the treatment effect is adjusted for the difference in the two groups on the mean value of the covariate. The *F* statistic still follows a non-central *F* distribution, $F(1, N\text{-}3, \lambda_x)$. However, notice that the denominator degrees of freedom is one less than the case without the covariate. The new non-centrality parameter is:

$$\lambda_x = \frac{N\beta_1^2}{4\sigma_{|x}^2} = \frac{N\beta_1^2}{4\sigma^2(1-\rho_{xy}^2)}. \quad [4.11]$$

Note that the smaller the conditional variance, the larger the non-centrality parameter, and greater the power of the test for large sample sizes.

In standardized notation, the non-centrality parameter is written as:

$$\lambda_x = \frac{N\delta^{*2}}{4} = \frac{N\delta^2}{4(1-\rho_{xy}^2)} \quad [4.12]$$

where

$\delta^* = \frac{\beta_1}{\sqrt{\sigma_{|x}^2}} = \frac{\beta_1}{\sqrt{\sigma^2(1-\rho_{xy}^2)}}$, the conditional effect size. Using Equation 4.11, we can calculate the power of the test as a function of the proportion of explained variation in the outcome by the covariate, the standardardized effect size, and the total sample size.

4.8 Using the Optimal Design for single level trials

The single level trial module allows the researcher to approach the power calculations using either the power determination approach or the effect size approach. The module menu is below:

Power on y-axis

Power vs. total number of people (N)

Power vs. effect size (δ)

Power vs. explained variation by covariate (R^2)

MDES on y-axis

MDES vs. total number of people (N)

MDES vs. power (*P*)

MDES vs. explained variation by covariate (R^2)

The first three options present the power on the y-axis and the sample size, effect size, and explained variance on the x-axis, respectively. The second three options present the effect size on the y-axis and the sample size, power, and explained variance on the x-axis. We present an example below and go through the steps involved in conducting a power analysis for the example varying the known and unknown parameters.

4.9 Example

A team of researchers is planning to do an experiment to determine attending a charter school compared to the local public school improves academic achievement. Assume that more students apply for admission to the charter school than they can admit. Because of the large number of applicants, all students enter a lottery and half of the students are randomly chosen to receive the treatment, enrollment at the charter school, and half of the students will receive the control condition, enrollment at the local public school. The researchers hypothesize that students enrolled in the charter school will have greater achievement than the students at the regular public schools. They plan to measure achievement using the Iowa Test of Basic Skills (ITBS). Section 4.10 presents a scenario in which the power determination approach for conducting a power anlaysis is most applicable to this study and provides the details of how to do the power analysis using OD. Section 4.11 presents a scenario in which the effect size approach for conducting a power anlaysis is most applicable to this study and provides the details of how to do the power analysis using OD.

4.10 Power determination approach for conducting a power analysis

Based on pilot study results, the researchers expect that students in the treatment group will score 0.25 standard deviation units greater than students in the control group on the ITBS. The researchers want to be able to detect this size treatment effect with power = 0.80. How many students are required for the study? Suppose the researchers decide to administer a pre-test to all students prior to the study. Based on past literature, they expect the pre-test to explain 64% of the variation in the post-test scores. How many students are required after including the pre-test in the design and analysis plan?

In this scenario, the total number of individuals is unknown and the effect size for planning is set at 0.25. Thus the most appropriate choice for the power analysis is to allow the sample size to vary on the x-axis and the power to vary on the y-axis. The steps follow.
Step 1: Select Person randomized trials → single level trials → Power on y-axis → power vs. total number of people (*N*). The blank screen is in Figure 4.1.

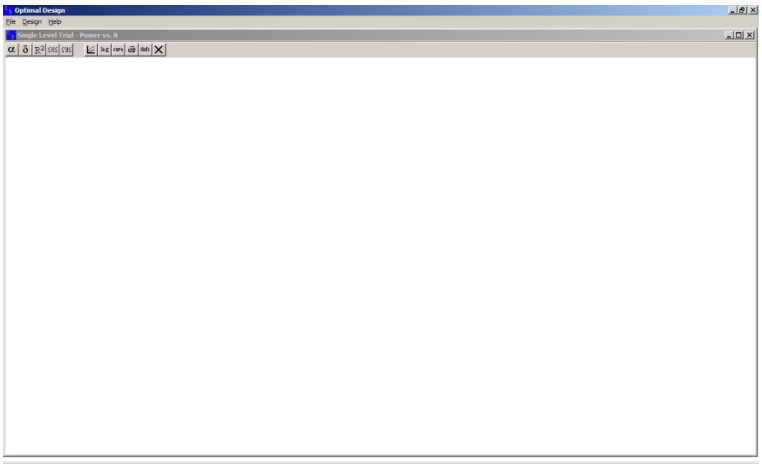


Figure 4.1. Main menu for a cluster randomized trial.

Note that the two parameters on the toolbar that are required for calculating the power include δ the effect size, and R^2, the percent of variation explained by the covariate (if there is a covariate).

Step 2: Click on δ. Set delta(1) = 0.25. The power curve appears in Figure 4.2.

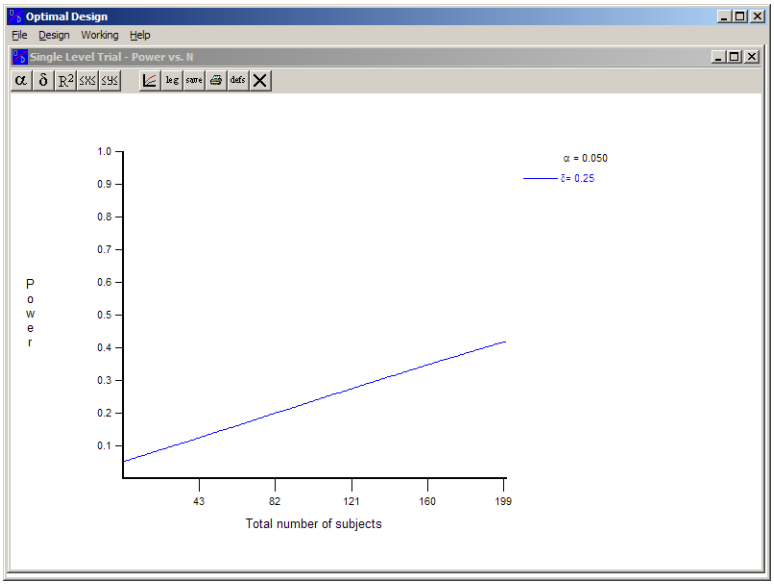


Figure 4.2. Power curve.

Step 3: Looking at the graph, we can see that we need to extend the x-axis in order to determine how many individuals are required to achieve power = 0.80. Click on <x< and set the maximum = 600. Figure 4.3 displays the screen.

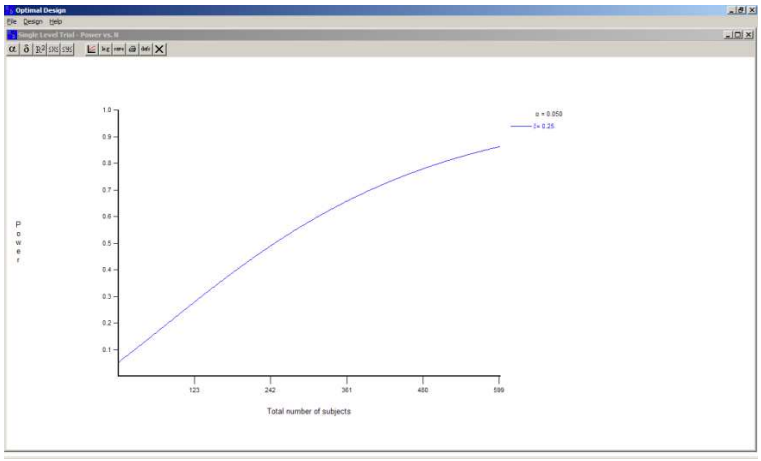


Figure 4.3. Power vs. total number of subjects.

Note that the key in the upper right corner shows the δ=0.25 that we specified. Clicking along the trajectory reveals that 504 people are required to detect an effect size of 0.25 with power = 0.80. This means 252 individuals would be randomized to both treatment and control.

Note that Figure 4.3 does not use the information in the covariate. Let's include the covariate and see what happens to the required sample size.

Step 4: Click on R^2. Set r2(2) = 0.64. Figure 4.4 displays the result.

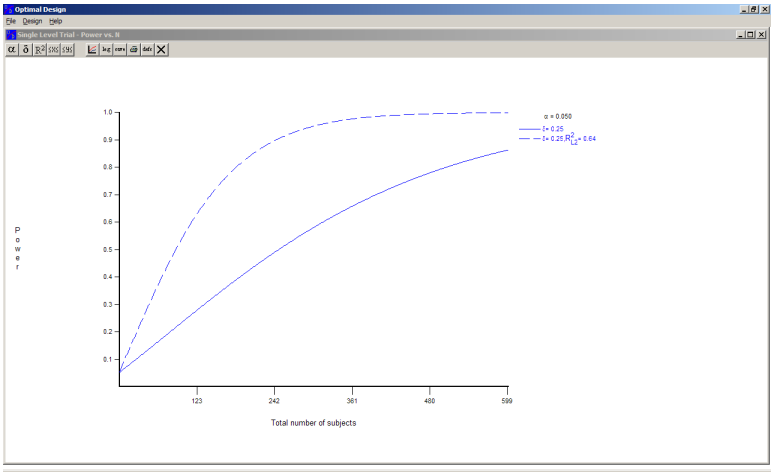


Figure 4.4. Power vs. total number of subjects with covariate.

The key indicates that the dotted trajectory represents the plot for the design with the pre-test. Clicking along the trajectory, we can see that the required sample size for power = 0.80 drops to 180, or 90 in each condition. Including the covariate reduces the total sample size by 324 persons. This reduction may be critical for reducing the cost of the experiment.

In this scenario, we allowed the sample size to vary along the x-axis. However, we could also choose to allow the effect size to vary along the x-axis (Power vs. effect size) or the explained variation by a covariate (Power vs. explained variation by a covariate) to vary along the x-axis and still maintaining the power on the y-axis.

4.11 Effect size approach for conducting a power analysis

Suppose that the researchers have counted up the total number of people that entered the lottery and discover that there are 200 people that want to participate. All 200 people will enter the lottery, thus 100 people will be assigned to the treatment and 100 people will be assigned to the control. What is the minimum detectable effect size (MDES) the researchers can find with power = 0.80? Suppose the researchers decide to administer a pre-test to all kids prior to the study. Based on past literature, they expect the pre-test to explain 64% of the variation in the post-test scores. What is the MDES with power = 0.80?

In Scenario 2, the MDES is unknown and the total sample size is limited to 200. Thus the most logical approach for conducting the power analysis is to allow the MDES to vary on the y-axis. One option then is to select the following:

Step 1: Select Person randomized trials → single level trials → MDES on y-axis → MDES vs. number of people (*N*). The blank screen appears in Figure 4.5.

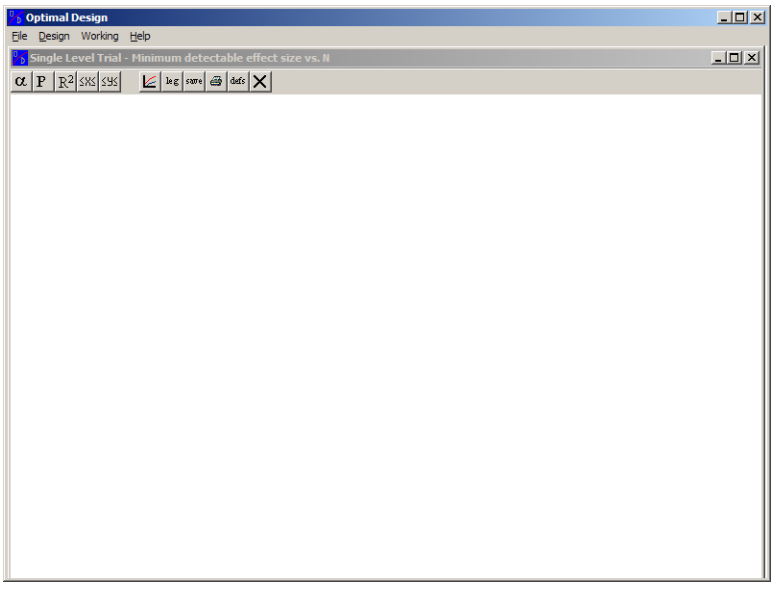


Figure 4.5. Main menu for person randomized trial.

The toolbar is identical to the toolbar in Figure 4.5 except for the required design parameters. Because the MDES is on the y-axis and the sample size is on the x-axis, the program

requires that the user specify power and R^2, the percent of variation explained by the covariate. To determine the MDES for power of 0.80, follow the steps below:

Step 1: Click on *P*. Set *P*(1) = 0.80.

Step 2: Click on R^2. Set r2(2) = 0.64. Figure 4.6 displays the results.

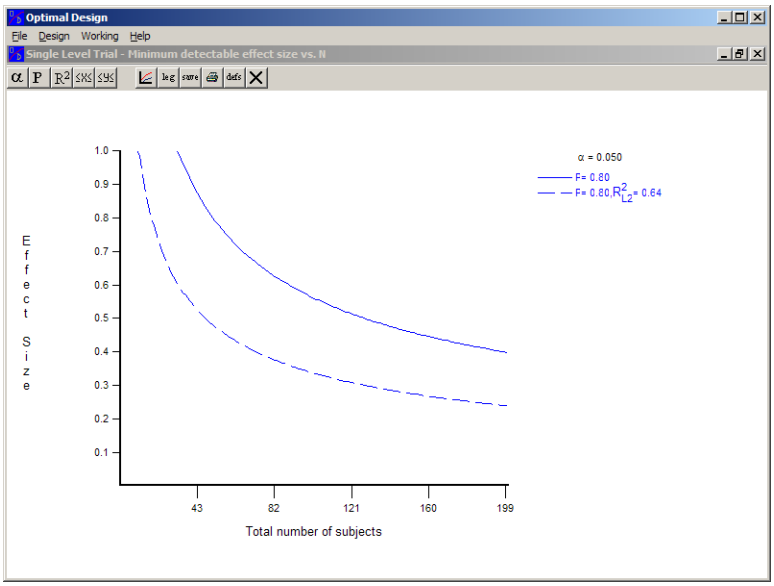


Figure 4.6. MDES vs. power.

Clicking along the solid trajectory reveals a MDES of 0.40 for 200 people whereas clicking along the dotted trajectory reveals a MDES of 0.24 for 200 people, both under the constraint of power = 0.80.

In this scenario, we allowed the total of people to vary on the x-axis. Additionally, the power (MDES vs. power) or explained variation by a covariate (MDES vs. explained variation by a covariate) could vary along the x-axis with MDES on the y-axis.

5.0 Multi-site (Blocked) trials

We define a multisite or blocked trial as a two-level design with students within blocks. For example, classrooms may represent a block and within each classroom, students are randomly assigned to receive a novel treatment. We consider power for the treatment effect, first assuming random site effects after which we consider fixed site effects.

5.1 The model (Assuming random site effects)

The model for a multi-site trial can be thought of as a two level hierarchical linear model. The level-1, or individual level model is:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad [5.1]$$

for

$i = 1...n$ persons per site

$j = 1,..., J$ sites

where

Y_{ij} is the response for person i at site j

β_{0j} is the mean response at site j

β_{1j} is the treatment effect at site j

X_{ij} is the treatment indicator with ½ for treatment and -½ for control

r_{ij} is the random error associated with person i at site j

σ^2 is the between persons variation.

The level-2, or site level model is:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} & u_{0j} &\sim N(0, \tau_{00}) \\ \beta_{1j} &= \gamma_{10} + u_{1j} & u_{1j} &\sim N(0, \tau_{11}) \end{aligned} \qquad [5.2]$$

where

γ_{00} is the grand mean

γ_{10} is the main effect of treatment

u_{0j} is the random error associated with the mean

u_{1j} is the random error associated with the treatment effect

τ_{00} is the variability between site means

τ_{11} is the variability between sites on the treatment effect

The inclusion of the random error terms, u_{0j} and u_{1j}, is what defines this as a random effects model. Our primary interest is the main effect of treatment, γ_{10}, and the variability of the treatment effect across sites, τ_{11}.

5.2 Testing the treatment effect

In a balanced design, the main effect of treatment is estimated by:

$$\hat{\gamma}_{10} = \bar{Y}_E - \bar{Y}_C \qquad [5.3]$$

where

$\bar{Y}_E$ is the mean for the experimental group

$\bar{Y}_C$ is the mean for the control group.

The variance of the estimated treatment effect is (Raudenbush & Liu, 2000):

$$\text{var}(\hat{\gamma}_{10}) = \frac{\tau_{11} + 4\sigma^2 / n}{J}. \qquad [5.4]$$

Note that the variance of the estimated treatment effect is a function of the number of blocks, *J*, the number of persons per block, *n*, the between-persons variation, σ^2, and the variability between sites on the treatment effect, τ_{11}.

If the data are balanced, we can use the results of an analysis of variance with random effects for the sites and fixed effects for the treatment. The *F* statistic for testing the main effect of treatment follows a non-central *F* distribution, $F(1, J\text{-}1, \lambda)$. Recall that the non-centrality parameter, λ, is the ratio of the squared treatment effect to the variance of the treatment effect estimate. The non-centrality parameter can be written as:

$$\lambda = \frac{J\gamma_{10}^2}{\tau_{11} + 4\sigma^2 / n} \qquad . \qquad [5.5]$$

Recall the larger the non-centrality parameter, the greater the power. It is clear that increasing the number of sites as well as the number of persons per site increases λ. However, looking at Equation 5.5 we can see that *J* is more influential for increasing λ than is *n*. In addition, studies

attempting to detect larger effect sizes have greater power. Finally, as the treatment effect variability gets larger, *J* becomes exceedingly important. In cases with extremely large effect size variability and small treatment effects, it is important to recognize that the treatment effect may not be very meaningful. For example, a large effect size variability may mean that in some sites, the treatment is producing a harmful, or negative effect. The average effect may be positive but it may be hiding the fact that the treatment works very well in some sites and is harmful in other sites. Hence it is important to report both the estimate of the treatment effect and the variability in the treatment effect in the results of a multi-site trial.

5.3 Standardized notation

In order to give meaning to the size of an effect without knowledge of the specific outcome scale or measurement, we often standardize the effect sizes. In a multi-site trial, we also need to standardize the effect size variability. For example, an effect size variance of 0.10 is the same as a standard deviation of approximately $\sqrt{0.10} = 0.31$. If a researcher desires a minimum detectable effect of 0.20, a standard deviation of 0.31 is large and would indicate a lot of variability in the treatment effect across sites. Indeed, if the treatment effects were normally distributed, we would expect 95% of them to lie within about two standard deviations of the mean; more precisely, in the interval $.10 \pm 1.96 * 0.31 = (-0.51, 0.71)$, indicating that the effect can range from very harmful to very positive.

Dividing the numerator and denominator of Equation 5.5 by σ^2, we can express the non-centrality parameter as:

$$\lambda = \frac{J\gamma_{10}^2 / \sigma^2}{(\tau_{11} + 4\sigma^2 / n) / \sigma^2} = \frac{J\delta^2}{\sigma_\delta^2 + 4/n} \qquad [5.6]$$

where

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma^2}}, \quad \sigma_\delta^2 = \frac{\tau_{11}}{\sigma^2} \qquad [5.7]$$

are, respectively, the standardized effect size and the standardized effect size variability. Note that 5.6 shows that the power in fact depends only on *n, J,* δ, and σ_δ^2.[2]

5.4 The model with a covariate

The covariate reduces the between-person variation, hence increasing the precision of the estimate of the treatment effect. The proportion of variance explained by the covariate is denoted ρ_{xy}^2. Level-1 of the model includes the covariate:

$$Y_{ij} = \beta_{0j} + \beta_{1j} W_{ij} + \beta_{2j} X_{ij} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2_{|x}) \qquad [5.8]$$

Note: $\sigma^2_{|x} = (1 - \rho_{xy}^2)\sigma^2$ for

$i = 1...n$ persons per site

$j = 1,..., J$ sites

where

Y_{ij} is the response for person i at site j

β_{0j} is the adjusted mean response at site j

β_{1j} is the adjusted treatment effect at site j

W_{ij} is the treatment indicator with ½ for treatment and -½ for control

X_{ij} is the covariate

e_{ij} is the random error associated with person i at site j

σ^2 is the between persons variation conditional on the covariate.

The level-2, or site level model is:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \qquad & u_{0j} &\sim N(0, \tau_{00|x}) \\ \beta_{1j} &= \gamma_{10} + u_{1j} \qquad & u_{1j} &\sim N(0, \tau_{11}) \\ \beta_{2j} &= \gamma_{20} \end{aligned} \qquad [5.9]$$

---

[2] The OD software is based on parameter estimates prior to blocking as well as an estimate of the percent of variance explained by the blocking variable. After the user enters the parameters, the program calculates the parameters defined in equation 7 as follows: $\delta = \frac{\delta_u}{\sqrt{(1-B)}}$ and $\sigma_\delta^2 = \frac{\sigma_{\delta_u}^2}{(1-B)}$ where *u* is the value prior to blocking and B is the percent of variance explained by blocking.

where

γ_{00} is the grand mean

γ_{10} is the average treatment effect

γ_{20} is the regression coefficient for the cluster-level covariate, which is assumed constant across sites

u_{0j} is the random effect associated with the mean

u_{1j} is the random effect associated with the treatment effect

τ_{00} is the residual variance between sites means

τ_{11} is the variability between sites on the treatment effect.

The inclusion of the random error terms, u_{0j} and u_{1j}, is what defines this as a random effects model. Our primary interest is the main effect of treatment, γ_{10}, and the variability of the treatment effect, τ_{11}.

5.5 Testing the treatment effect (including a covariate)

The estimate of the treatment effect is:

$$\hat{\gamma}_{10} = \bar{Y}_E - \bar{Y}_C - \hat{\gamma}_{20}(\bar{X}_E - \bar{X}_C). \qquad [5.10]$$

The estimate is adjusted for the difference in the two groups on the mean value of the covariate. The variance of the treatment effect estimate is:

$$\text{var}(\hat{\gamma}_{10}) = \frac{\tau_{11} + 4\sigma^2_{|x}/n}{J}. \qquad [5.11]$$

where

$\sigma^2_{|x}$ is the conditional variance, $(1-\rho^2_{xy})\sigma^2$.

The *F* statistic still follows a non-central *F* distribution, $F(1, J\text{-}1, \lambda_x)$. The new non-centrality parameter is:

$$\lambda_x = \frac{J\gamma^2_{10}}{\tau_{11} + 4\sigma^2_{|x}/n} \qquad [5.12]$$

Note that the smaller the conditional variance, the larger the non-centrality parameter, and greater the power of the test for large sample sizes.

In standardized notation, the noncentrality parameter is written as:

$$\lambda_x = \frac{J\delta^{*2}}{\sigma_\delta^{*2} + 4/n}$$

where

$\delta^* = \frac{\gamma_{10}}{\sqrt{\sigma^2_{|x}}} = \frac{\delta}{\sqrt{1-\rho^2_{xy}}}$, the conditional effect size and

$\sigma_\delta^{2*} = \frac{\sigma_\delta^2}{1-\rho^2_{xy}}$.

5.6 Testing the variance of the treatment effect

For any design with $J \geq 4$, we can estimate and test the variance of the treatment effect across sites. This is particularly important if the treatment effect variability is non-negligible, in which case the main effect of treatment may poorly represent the treatment effect in any specific site. In this case, we need to have adequate power to detect this variability.

The power to detect the variance of the treatment effect is also based on an *F*-test. In standardized notation, the *F*-statistic is (Raudenbush and Liu, 2000)

$$F = \frac{n\hat{\tau}_{11} + 4\hat{\sigma}^2}{4\hat{\sigma}^2}. \qquad [5.13]$$

The *F*-statistic follows a central *F* distribution with *J-1, J(n-2)* numerator and denominator degrees of freedom. The ratio of the expectation of the numerator to the denominator is

$$\omega = \frac{n\tau_{11} + 4\sigma^2}{4\sigma^2} = \frac{n\sigma_\delta^2 + 4}{4} = 1 + \frac{n\sigma_\delta^2}{4}. \qquad [5.14]$$

Under the null hypothesis of no effect size variability, we expect σ_δ^2 to be 0, thus ω=1. As the ratio of expected mean squares increases, so does the power to detect the effect size variability. We can see from equation 9 that as σ_δ^2 or *n* get larger, $\frac{n\sigma_\delta^2}{4}$ also gets larger, which means the power increases. This contradicts what we learned about increasing the power of the test to detect the main effect of treatment. For that test, increasing the number of sites yields greater increases in power than the number of individuals per sites, and smaller variance in the treatment effect across sites results in larger power. Thus studies cannot be planned to maximize the power to detect the main effect of treatment and the variance of the treatment effect simultaneously. Prior to planning a study, researchers must decide the primary goal of the study, detecting the

main effect of treatment or the magnitude of the treatment by site variance, and plan the study accordingly.

5.7 The fixed effects model

The fixed effects model assumes homogeneity of the treatment effect across sites. The fixed effects model looks the same as the random effects model except that u_{0j} and u_{1j} are designated as fixed constants rather than random variables. This difference is depicted in the level-2 model:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}\end{aligned} \qquad [5.15]$$

where

γ_{00} is the grand mean

γ_{10} is the main effect of treatment

u_{0j} $j=1,\ldots,J$ are fixed effects associated with each site mean, and are constrained to have a mean of zero.

u_{1j} $j=1,\ldots,J$ are fixed effects associated with each site treatment effect, and are constrained to have a mean of zero.

We are interested in the main effect of treatment, γ_{10}, and the fixed treatment by site interaction effects, u_{1j}, $j=1,\ldots,J$.

5.8 Testing the treatment effect

We can use the results from an analysis of variance with fixed effects for the site and the main effect of treatment as well as the fixed effects. Again, the test statistic is an F statistic. The F test follows a noncentral F distribution, F(J-1, J(n-2); λ)**.** In standardized notation, the noncentrality parameter is:

$$\lambda = \frac{Jn\delta^2}{4} = \frac{J\delta^2}{4/n}. \qquad [5.16]$$

Note that the treatment effect variability does not appear in the formula because we do not allow the treatment effect to vary randomly across sites. The models and non-centrality parameters are easily extended to the case with a covariate by following the logic presented in sections 5.4 and 5.5. However, we caution that the main effect of treatment will be uninteresting or even

misleading when effect size variability is large, that is when u_{1j} $j=1,\ldots,J$ vary substantially. In that case, the main effect may poorly represent the treatment effect in any given site, and one would want to estimate and test $j=1,\ldots,J$. The tests for these specific site-by-site treatment effects (see Section 5.9) may be poor, particularly when the sample size per site is small

5.9 Testing site-by-treatment variation in the context of a fixed effects model

Operationally, the test of the treatment by site variation for a fixed effects model is the same as that for a random effects model. The primary difference is in the null hypothesis. In the random effects model, we test:

$$H_0 : \sigma_\delta^2 = 0 . \qquad [5.17]$$

However, in a fixed effects model, the treatment by site effects are fixed constants so we test:

$$H_0 : \sum_{j-1}^{J} u_{1j}^2 = 0 . \qquad [5.18]$$

We use the same *F* statistic, $F = \frac{MS_{trmtxsite}}{MS_{withincell}}$ with *J-1* numerator degrees of freedom and *J(n-2)* denominator degrees of freedom. If we reject the null hypothesis, a logical next step would be to try to identify sites for which the treatment effect is the same (Kirk, 1982).

5.10 Using the Optimal Design for multisite (blocked) trials

The multisite (blocked) trial module allows the researcher to approach the power calculations using either the power determination approach or the effect size approach. The module menu is below:

Power for treatment effect on y-axis

- Power vs. site size (n)
- Power vs. total number of sites (*J*)
- Power vs. effect size (δ)
- Power vs. effect size variability

MDES on y-axis

- MDES vs. site size (n)
- MDES vs. total number of sites (*J*)
- MDES vs. effect size variability
- MDES vs. power (*P*)

Power for effect size variability on y-axis

Power vs. site size (n)

Power vs. total number of sites (*J*)

We present an example below and go through the steps involved in conducting a power analysis for the example varying the known and unknown parameters.

5.11 Example

Suppose a team of researchers is planning to test a new tutoring program for at risk 2nd grader students in a particular school district. At-risk 2nd graders in the district will be randomly assigned to either the treatment condition, a new pullout tutoring program, or the control condition, the standard in class tutoring program. Researchers plan to block on classroom. Thus within each classroom the identified as-risk 2nd graders will be assigned to the treatment or the control condition. The researchers expect that blocking on classroom will explain 30% of the variation in the outcome. The researchers plan to use a random effects model and assume the effect size variability to be 0.01. Section 5.12 presents a scenario in which the power determination approach for conducting a power analysis is most applicable to the study and provides the details of how to do the power analysis using OD. Section 5.13 presents a scenario in which the effect size approach for conducting a power analysis is most applicable to the study and provides the details of how to do the power analysis using OD.

5.12 Power determination approach for conducting a power analysis

Based on pilot study results, the researchers expect that students in the treatment group will score 0.25 standard deviation units greater than students in the control group on the outcome. The researchers want to be able to detect this size treatment effect with power = 0.80. They have 20 students per classroom. How many classrooms are required for the study? Suppose the researchers decide to administer a pre-test to all students prior to the study. Based on past literature, they expect the pre-test to explain 50% of the variation in the post-test scores. How many classrooms are required after including the pre-test in the design and analysis plan?

In this scenario, the total number of classrooms is unknown and the effect size for planning is set at 0.25. Thus the most appropriate choice for the power analysis is to allow the sample size to vary on the x-axis and the power to vary on the y-axis. The steps follow.

Step 1: Select Person randomized trials → multisite (blocked) trials → Power on y-axis → power vs. total number of sites (*J*). The blank screen is in Figure 5.1.

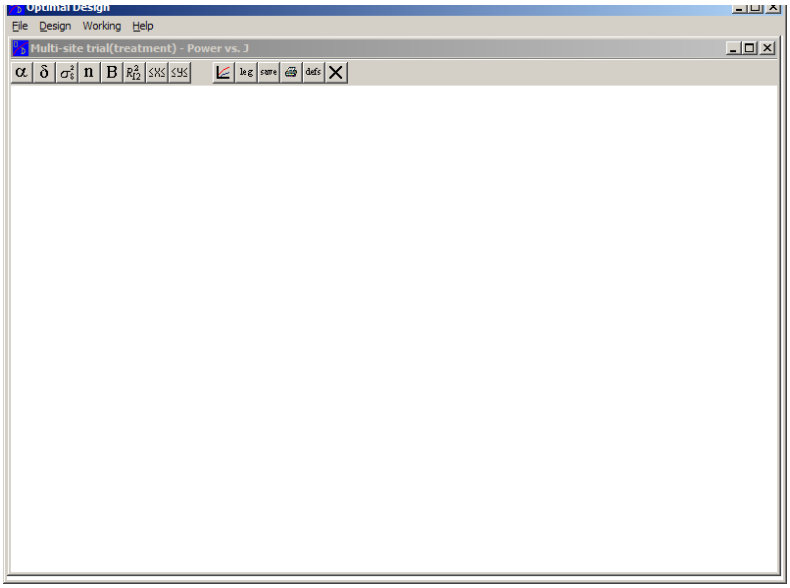


Figure 5.1. Main menu for person randomized trials with blocking.

Step 2: Click on δ. Set delta(1) = 0.25.

Step 3: Click on σ_δ^2. Set $\sigma_\delta^2 = 0.01$.

Step 4: Click on n. Set n(1) = 20.

Step 5: Click on B. Set B(1) = 0.30. The power curve appears in Figure 5.2.

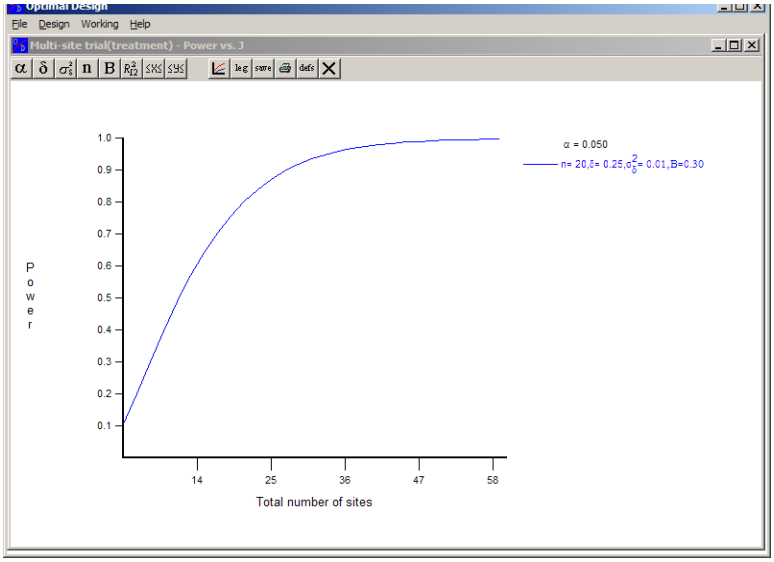


Figure 5.2. Power vs. total number of sites.

Note that the key in the upper right corner shows the δ=0.25 that we specified. Clicking along the trajectory reveals that 21 sites or classrooms are required to detect an effect size of 0.25 with power = 0.80. Note that Figure 5.2 does not account for the covariate. Let's include the covariate and see what happens to the required sample size.

Step 6: Click on R^2. Set r2(2) = 0.50. Figure 5.3 displays the result.

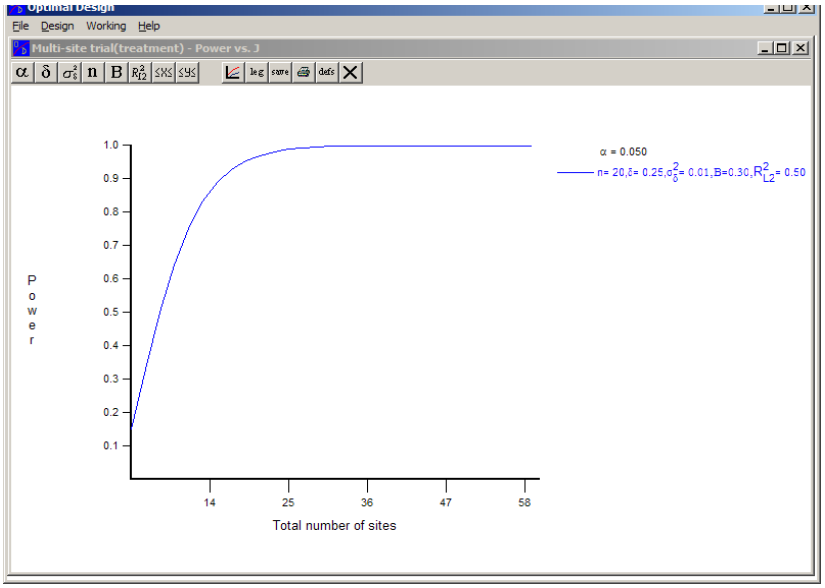


Figure 5.3. Power vs. total number of subjects.

Clicking along the trajectory, we can see that the required sample size for power = 0.80 drops to 13 sites, assuming 20 individuals per sites. This reduction may be critical for reducing the cost of the experiment.

In this scenario, we assumed a random effects model. We could easily change it to a fixed effects model by setting the effect size variability to 0. However, it is critical to think about the implications of choosing fixed or random site effects from a practical perspective, and not a purely statistical power perspective.

5.13 Effect size approach for conducting a power analysis

Suppose that the researchers are limited to 20 classrooms with 20 individuals per classroom. They are still interested in an effect size of 0.25. What is the minimum detectable effect size (MDES) the researchers can find with power = 0.80? Suppose the researchers decide to administer a pre-test to all kids prior to the study. Based on past literature, they expect the pre-test to explain 50% of the variation in the post-test scores. What is the MDES with power = 0.80?

In Scenario 2, the MDES is unknown. Thus the most logical approach for conducting the power analysis is to allow the MDES to vary on the y-axis. One option then is to select the following:

Step 1: Select Person randomized trials → multisite (blocked) trials → MDES on y-axis → MDES vs. number of clusters (*J*). The blank screen appears in Figure 5.4.

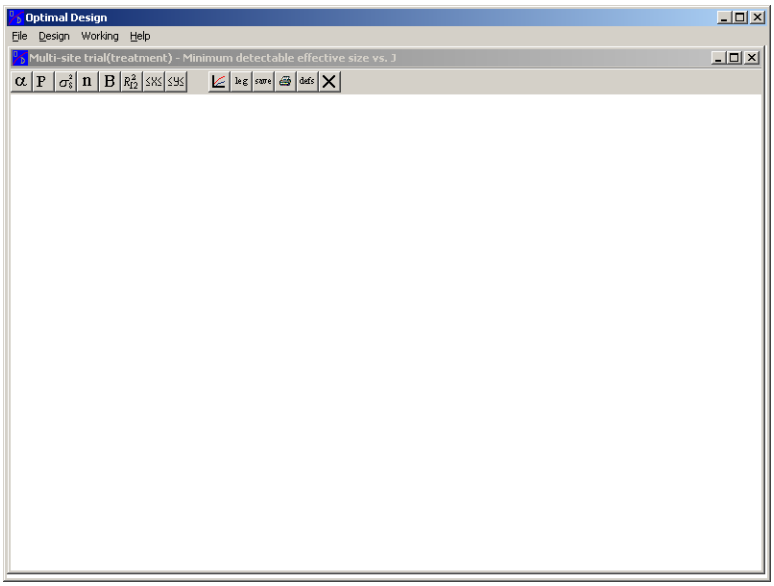


Figure 5.4. MDES vs. number of cluster (*J*).

The toolbar is identical to the toolbar in Figure 5.4 except for the required design parameters. Because the MDES is on the y-axis and the power is on the x-axis, the program requires that the user specify *J*, the total number of sites, n, the number of individuals per site, σ_δ^2 the effect size variability, B, the percent of variance explained by blocking, and R^2, the percent of variation explained by the covariate. To determine the MDES for power of 0.80, follow the steps below:

Step 1:Click on *P*. St *P*(1) = 0.80.

Step 2: Click on σ_δ^2. Set σ_δ^2 = 0.01.

Step 3: Click on n. Set n(1) = 20.

Step 4: Click on B. Set B(1) = 0.30. Figure 5.5 displays the results.

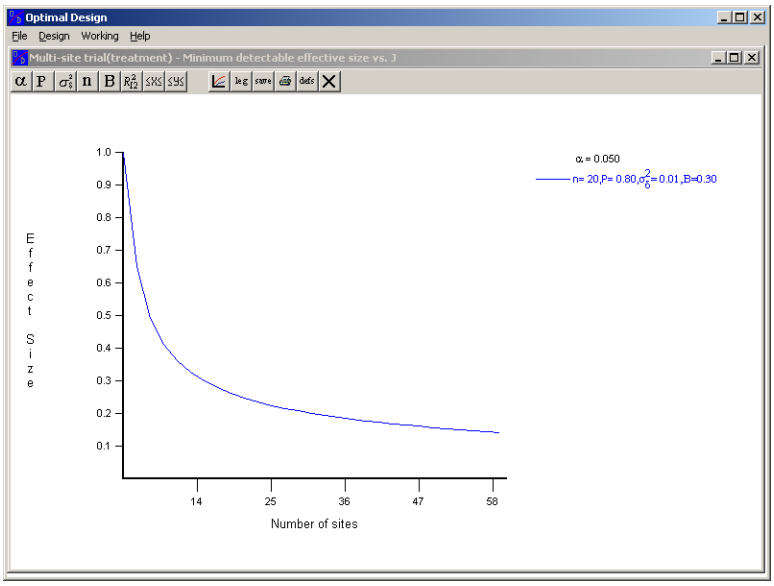


Figure 5.5. MDES vs. number of clusters (*J*).

Clicking along the trajectory reveals a MDES of approximately 0.26 with $J = 20$. Next we can add the covariate.

Step 5: Click on R^2. Set $R^2 = 0.50$. Figure 5.6 displays the results.

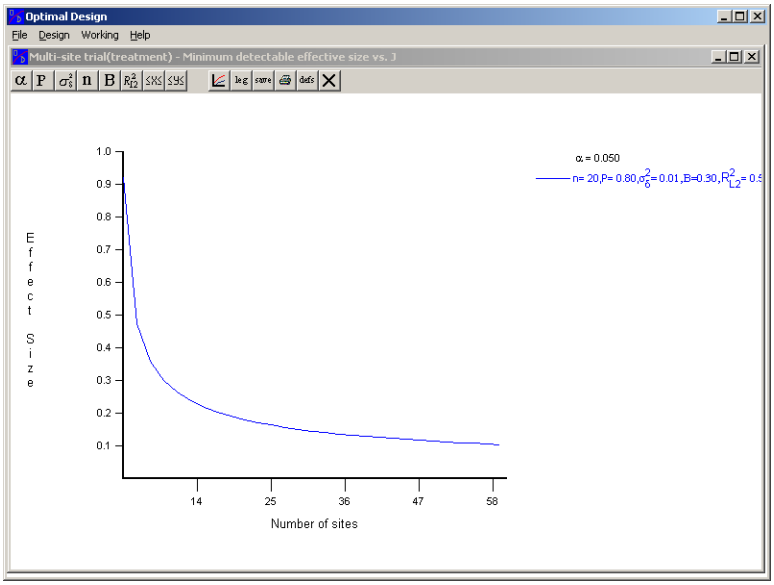


Figure 5.6. MDES vs. power with a covariate.

Clicking along the trajectory reveals an effect size of about 0.19. We assumed random site effect but again could change this by setting the effect size variability to 0.

6.0 Repeated measures trials

Similar to single level trials, repeated measure trials rely on the assignment of individuals to treatments. However, in a trial with repeated measures, individuals are typically assessed prior to the treatment and then multiple times after the treatment is implemented. By tracking individuals over time, researchers are able to assess the group effects on individual growth.

The general format of a repeated measure trial is as follows: 1) randomly assign individuals to treatment or control, 2) assess students in the treatment and control group prior to implementation of the treatment, 3) implement the treatment for the treatment group, 4) assess the students in both groups on the outcome of interest, 5) repeat assessments of students in both groups a pre-determined number of times over equally spaced time intervals. By collecting repeated measures on individuals, we are able to model individual growth trajectories. We can model linear or curvilinear trajectories. A linear trajectory, or first degree polynomial, is characterized by an intercept and a linear rate of change, or slope. If non-linear growth is expected, second, third, or higher degree polynomials may be added in order to model curvilinear trajectories. A second degree polynomial, also known as a quadratic polynomial, adds an acceleration parameter to the intercept and rate of change. A third degree polynomial, or a cubic polynomial, is characterized by four parameters, change in acceleration, rate of acceleration, linear rate of change, and an intercept. Individual growth trajectories are plotted in order to assess the average treatment effect on a specific polynomial change parameter.

The power for a design with repeated measures is more complicated than for a single level trial. To simplify the design calculations, we impose the following constraints: orthogonal designs, continuous outcomes, a linear link function, random effects covariance structure, homogeneous covariance structure within each treatment, and complete data. Fist we examine the statistical models.

6.1 The model

We can represent the data from a single level trial with repeated measures as a two-level hierarchical model, with occasions nested within persons. The general level one model for a polynomial change parameter of order *p* is:

$$Y_{mi} = \sum_{p=0}^{p-1} \pi_{pi} c_{pm} + e_{mi}, \quad e_{mi} \sim N(0, \sigma^2). \qquad [6.1]$$

for

i where *i*=1,…,*N p*ersons

m where *m=1,…M* time points

where

p indicates the polynomial order of change (ie. linear, quadratic, cubic)

c_{pm} is the orthogonal polynomial contrast coefficient

π_{pi} is the level one coefficient of polynomial order *p*

e_{mi} is the random error associated with the repeated measures

σ^2 is the level-1 variability, or measurement error.

The purpose of the polynomial contrast coefficients is to center the data, which makes the interpretation easier. The formulas for calculating the contrast coefficients are given below (Raudenbush and Liu, 2001):

$$
\begin{aligned}
c_{0m} &= 1 \\
c_{1m} &= m - \sum_{m=1}^{M} m / M \\
c_{2m} &= \frac{1}{2}\left(c_{1m}^2 - \sum_{m=1}^{M} c_{1m}^2 / M \right) \\
c_{3m} &= \frac{1}{6}\left(c_{1m}^3 - \frac{\sum_{m=1}^{M} c_{1m}^4}{\sum_{m=1}^{M} c_{1m}^2} c_{1m} \right)
\end{aligned}
\qquad [6.2]
$$

The general level two model is:

$$\pi_{pi} = \beta_{p0} + \beta_{p1} X_i + u_{pi} \qquad u_{pi} \sim N(0, \tau_{\pi p}) \qquad [6.3]$$

where

β_{p0} is the mean for the p[th] order polynomial change parameter

β_{p1} is the treatment effect for the p[th] order polynomial change parameter

X_i is an indicator for the treatment or control group, ½ for treatment, -½ for control

u_{pi} is the random effect associated with each person

$\tau_{\pi p}$ is the between-person variance for the p[th] order polynomial change parameter.

To illustrate, let us consider a 1st order polynomial change parameter, or linear model. The level-1 model is:

$$Y_{mi} = \pi_{0i} + \pi_{1i} c_{1m} + e_{mi} \qquad e_{mi} \sim N(0, \sigma^2) \qquad [6.4]$$

for

$m = 1,\ldots,M$ occasions

$i = 1, \ldots, I$ persons

where

π_{0i} is the mean response for person *i*

π_{1i} is the average rate of change for person *i*

$c_{1m} = m - \sum_{m=1}^{M} m / M = m - \overline{m}$ is the orthogonal linear contrast coefficient

e_{mi} is the measurement error

σ^2 is the within-person variability.

We can calculate the linear contrast coefficients for any M using equation 2. For example, if the total number of data points is 5, that is *M*=5, the orthogonal contrast coefficients for a first degree polynomial are:

$$c_0 = (1,1,1,1,1)$$
$$c_1 = (-2,-1,0,1,2). \qquad [6.5]$$

The level-2 model is:

$$\pi_{0i} = \beta_{00} + \beta_{01} X_i + u_{0i} \qquad u_{0i} \sim N(0, \tau_{\pi 0})$$
$$\pi_{1i} = \beta_{10} + \beta_{11} X_i + u_{1i} \qquad u_{1i} \sim N(0, \tau_{\pi 1}) \qquad [6.6]$$

where

β_{00} is the mean response across persons

β_{01} is main effect of treatment for the means

β_{10} is the average growth rate across persons

β_{11} is the main effect of treatment for the growth rates

X_i is an indicator for the treatment or control group, ½ for treatment, -½ for control

u_{0i} is the random effect associated with the mean

u_{1i} is the random effect associated with the growth rates

$\tau_{\pi 0}$ is the between-person variance in means

$\tau_{\pi 1}$ is the between-person variance in growth rates.

In this case, our primary interest is β_{11}, the main effect of treatment for the growth rates, and $\tau_{\pi 1}$, the variability in growth rates across persons.

6.2 Testing the treatment effect

The average treatment effect for linear change in a balanced design is estimated by:

$$\hat{\beta}_{p1} = \frac{\sum_{i \in E} \hat{\pi}_{1i}}{n_E} - \frac{\sum_{i \in C} \hat{\pi}_{1i}}{n_C}. \qquad [6.7]$$

where $\hat{\pi}_{1i} = \frac{\sum_{m=1}^{M} c_{1m} Y_{mi}}{\sum_{m=1}^{M} c_{1m}^2}$ is the person-specific ordinary least squares estimator of the linear slope, and n_E and n_C are the sample sizes of the experimental and control groups.

To estimate the treatment effect, we average over occasions and persons. The variance of the estimated treatment effect for the p[th] polynomial order of change is (Raudenbush and Liu, 2001):

$$Var(\hat{\beta}_{p1}) = \frac{4(\tau_{\pi_1} + V_p)/n}{J} \qquad [6.8]$$

where

$$V_p = \frac{\sigma^2}{\sum_{m=1}^{M} c_{pm}^2} = \frac{\sigma^2 f^{2p} (m - p - 1)!}{K_p (M + p)!} \qquad [6.9]$$

where

f is the frequency of observation

D is the duration of the study

M is the total number of occasions where $M = Df + 1$

p is the polynomial order of change

K_p is a constant where $K_1 = 1/12$, $K_2 = 1/720$, $K_3 = 1/100{,}800$

σ^2 is the measurement error.

The term V_p denotes the conditional variance of the least squares estimate of each person's change parameter. Note that V_p is a function of the frequency and the duration of the study.

In the case of the linear change model, the variance of the estimate of the treatment effect is:

$$\text{var}(\hat{\beta}_{11}) = \frac{4(\tau_{\pi 1} + V_1)/n}{J} \qquad [6.10]$$

where

$$V_1 = \frac{\sigma^2}{\sum_{m=1}^{M} c_{1m}^2} = \frac{\sigma^2 f^2 (m-2)!}{(1/12)(M+1)!} \qquad [6.11]$$

.

The test statistic for the test that the treatment effect for the pth order polynomial equals zero is an *F* statistic. When the treatment effect is non-zero, the test statistic follows a non-central *F* distribution, $F(1, N\text{-}2; \lambda)$. As previously noted, the larger the non-centrality parameter, the greater the power of the test. The non-centrality parameter can be expressed as the ratio of the squared true treatment effect to the variance of the estimate of the treatment effect:

$$\lambda = \frac{\beta_{p1}^2}{Var(\hat{\beta}_{p1})} = \frac{n\beta_{p1}^2}{4(\tau_{\pi p} + V_p)} \qquad [6.12]$$

Beginning with the sample size, it is clear that increasing *n* increases λ, hence increasing the power of the test. The sample size is particularly important if the between-person variance is large. Looking at the variance components, we can see that small values of $\tau_{\pi p}$, or between-person variability, also increases the power. Intuitively this makes sense. If there is less variability between-persons, the estimate will be more precise and the power of the test is greater. We can also see that smaller values of V_p will increase the power. Recall that

$V_p = \frac{\sigma^2}{\sum_{m=1}^{M} c_{pm}^2} = \frac{\sigma^2 f^{2p} (m-p-1)!}{K_p (M+p)!}$. Decreasing the measurement error, σ^2, decreases V_p. Also, increasing the frequency of observations and the total number of observations can decrease V_p, particularly for higher order polynomials.

6.3 The standardized model

Similar to the previous designs, we standardize the model to facilitate a common language among researchers. The standardized effect size for a particular polynomial of interest is defined as:

$$\delta_p = \frac{\beta_{p1}}{\sqrt{\tau_{pp}}} \qquad [6.13]$$

where

β_{p1} is the group difference on the polynomial of interest

τ_{pp} is the population variance of the polynomial of interest.

Replacing equation 12 with the standardized parameter, the new noncentrality parameter can be expressed as:

$$\lambda = \frac{n\delta_p^2 \alpha_p}{4} \qquad [6.14]$$

where

α_p is the reliability of the least squares estimator $\hat{\pi}_p$ and

$$\alpha_p = \frac{Var(\pi_p)}{Var(\hat{\pi}_p)} = \frac{\tau_{\pi p}}{\tau_{\pi p} + V_p}.$$

The reliability is the ability with which a researcher can discriminate between people on their growth rate of the polynomial of interest using the least squares estimate. The reliability can be calculated using the HLM software.

6.4 Using the Optimal Design for repeated measures trials

The menu for the repeated measures is given below. The menu includes option for standardized or nonstandardized parameters.

Power on the y-axis (standardized)

Power for treatment on linear change

Power for treatment on quadratic change

Power for treatment on cubic change

Power on the y-axis (nonstandardized)

Power for treatment on linear change

Power for treatment on quadratic change

Power for treatment on cubic change

Each of the power options function similarly, so for illustration purposes, we will use the option for power for treatment on linear change.

6.5 Example

Recall that a team of researchers are planning to do an experiment to determine whether an intervention, enrollment at a charter school, improves academic achievement. Because of the large number of applicants for the school, all students enter a lottery and half of the students are randomly selected from the lottery and assigned to receive the treatment, enrollment at the charter school, while the other half of the students are enrolled in the local public school. The researchers hypothesize that students enrolled in the charter school will have greater achievement than the students at the regular public school. They plan to assess all the students prior to the study and then one time for the next five years. Based on data from a pilot study, they expect the level-1 variability to be 1.0 and the level-2 variability to be 0.10. They also expect students' academic growth to be linear. Section 6.6 presents a scenario in which the power determination approach for conducting a power analysis is the most relevant and provides the details for this approach.

6.6 Power determination approach for conducting a power analysis

Based on past research, the researchers expect a standardized effect size on the linear growth parameter of 0.25. That is, the difference in linear growth for students in the charter school compared to students in the control school is 0.25. How many students are required to detect an effect size of 0.25 with power of 0.80?

In this example, the total number of individuals is the unknown parameter and the effect size for planning is set at 0.25. Thus the most appropriate choice for the power analysis is to include power on the y-axis. The steps follow:

Step 1: Select Person randomized trials → repeated measures → Power on the y-axis → Power for treatment on linear change. The blank screen appears in Figure 6.1.

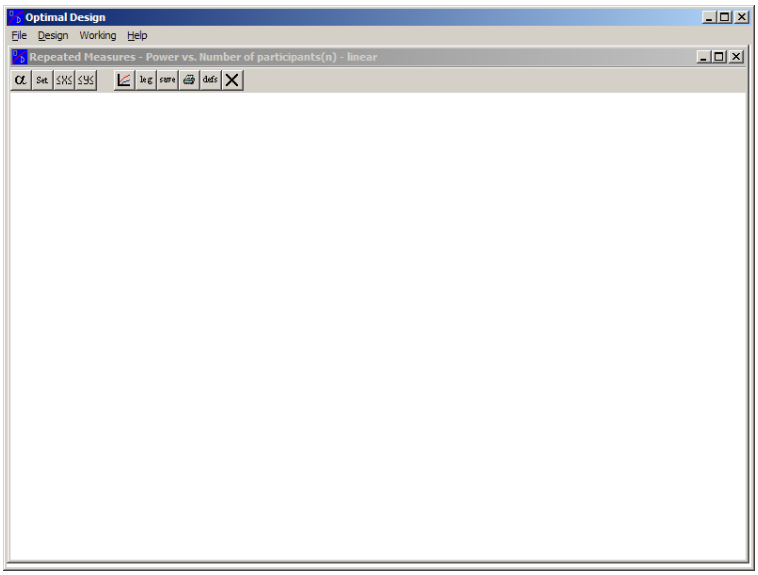
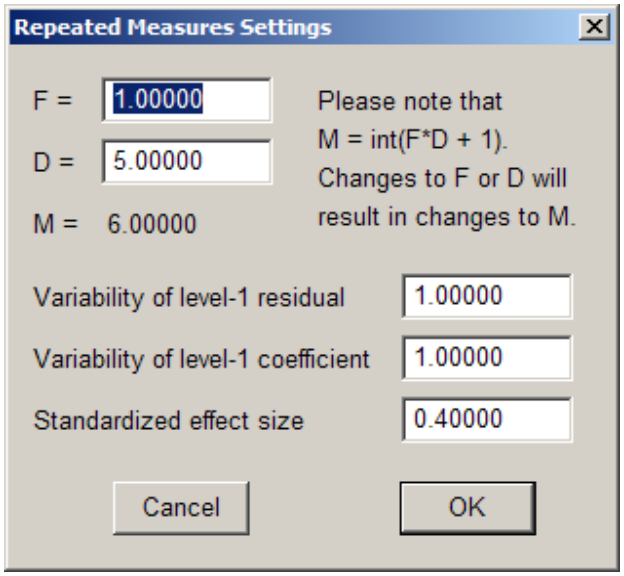


Figure 6.1. Blank screen for Power on y-axis → Power for treatment on linear change.

Step 2: Click on set. The set button brings up the screen in Figure 6.2.

Figure 6.2. The set button for linear change.



The following options appear within the set button:

F – specifies the frequency of the observation.

D – specifies the duration of the study.

M – is the total number of observations where M = fD+1. It is the product of the frequency times the duration plus the 1 observation that was pre-treatment.

Variability of level-1 residual – This is the measurement error, denoted σ^2 in the model.

Variability of level-1 coefficient – This is the between person variability on the polynomial of interest. For a linear growth model it is τ_{11} in the model.

Standardized effect size – This is δ, where $\delta = \frac{\beta_{11}}{\sqrt{\tau_{11}}}$ in the linear model.

After clicking on the set button, set F=1, D=5, M=6, variability of level-one residual = 1.0, variability of level-1 coefficient = 0.10, and standardized effect size = 0.25. We extend the x-axis to 800. Figure 6.3 displays the power curve.

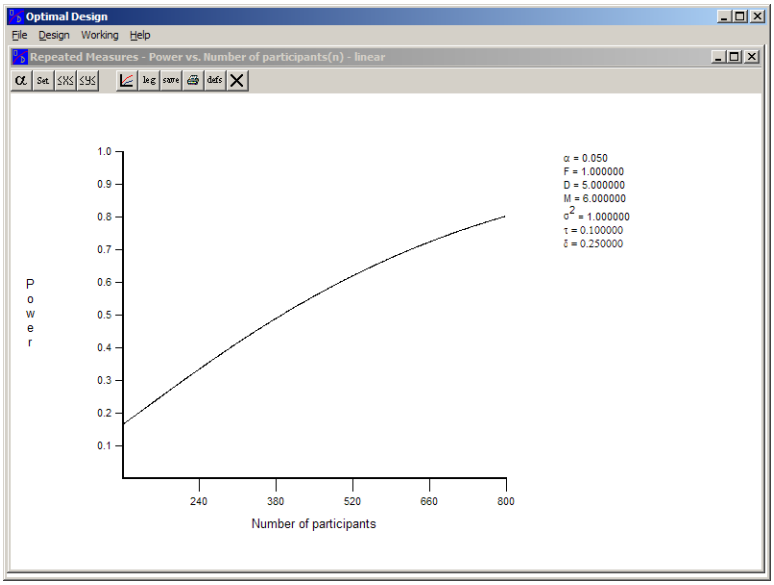


Figure 6.3. Power curve for repeated measures example.

Clicking along the trajectory, we can see that approximately 790 individuals are necessary to achieve power = 0.80. This is the same as 395 individuals per treatment condition.

There is also a function available for repeated measures trials that are non-standardized. It functions similarly to the example presented in this section. However, there is no option for MDES on the y-axis.

Section III: Optimal Design for cluster randomized trials

Optimal Design for cluster randomized trials includes trials where intact groups, or clusters, are randomly assigned to the treatment or control condition. For example, if students are nested within classrooms and classrooms are randomly assigned to either the treatment or control, the design is known as a two-level cluster randomized trial. There are fives designs in this category: three that do not include blocking and two that do include blocking. The non-blocked designs include the two-level cluster randomized trial (2-level CRT), the three level cluster randomized trials (3-level CRT), and the cluster randomized trial with repeated measures (CRT RM). The blocked designs include the three-level multi-site cluster randomized trials (3-level MSCRT) and the four-level multi-site cluster randomized trial (4-level MSCRT). We describe the conceptual details of each design and provide a "how to" guide for each design in the following 5 chapters.

7.0 Two-level cluster randomized trials

Two-level cluster randomized trials are studies in which individuals are nested within clusters and the clusters are randomly assigned to the treatment or control condition. For example, students are nested within classrooms and classrooms are randomly assigned to the treatment or control condition. For example, a team of researchers is interested in the effectiveness of a new math series. They decide to randomly assign schools to either the new series or the standard series. They plan to test students from one classroom within each school. In this case, schools are the unit of randomization and the students are nested within schools, making this a two-level cluster randomized trial. The power for a cluster randomized trial is more complicated than a single level trial since there is more than one level. We begin by examining the underlying statistical models.

7.1 The models

We can represent the data for a cluster randomized trial in hierarchical form, with individuals nested within clusters. The level-1, or person-level model is:

$$Y_{ij} = \beta_{0j} + e_{ij}, \qquad e_{ij} \sim N(0, \sigma^2) \qquad [7.1]$$

for $i \in \{1,2,...,n\}$ persons per cluster and $j \in \{1,2,...,J\}$ clusters,

where Y_{ij} is the outcome for person *i* in cluster *j*;

β_{0j} is the mean for cluster *j*;

e_{ij} is the error associated with each person; and

σ^2 is the within-cluster variance.

The level-2 model, or cluster-level model is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \qquad u_{0j} \sim N(0, \tau) \qquad [7.2]$$

where γ_{00} is the grand mean;

γ_{01} is the mean difference between the treatment and control group or the main effect of treatment;

W_j is the treatment contrast indicator, ½ for treatment and -½ for control;

u_{0j} is the random effect associated with each cluster; and

τ is the variance between clusters.

Replacing (2) in (1) yields the mixed model:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + e_{ij}, \quad u_{0j} \sim N(0, \tau) \text{ and } e_{ij} \sim N(0, \sigma^2). \qquad [7.3]$$

7.2 Testing the treatment effect

We are primarily interested in the main effect of treatment, γ_{01}, estimated by:

$$\hat{\gamma}_{01} = \bar{Y}_E - \bar{Y}_C, \qquad [7.4]$$

where $\bar{Y}_E$ is the mean for the experimental group and $\bar{Y}_C$ is the mean for the control group. When each treatment has an equal number, *J*/2, of clusters, the variance of the main effect of treatment is (Raudenbush, 1997):

$$Var(\hat{\gamma}_{01}) = \frac{4(\tau + \sigma^2 / n)}{J} \qquad [7.5]$$

where *n* is the total number of participants per cluster and *J* is the total number of clusters.

We can use hypothesis testing to determine if the main effect of treatment is "statistically significant," that is, not readily attributable to chance. Recall that a two-tailed null hypothesis states there is no difference whereas the alternative hypothesis states there is a difference. In symbols:

$$H_0 : \gamma_{01} = 0$$

$$H_1 : \gamma_{01} \neq 0$$

If the data are balanced, that is, there is an equal number of participants in each cluster, we can use the results of a two factor nested ANOVA to test the main effect of treatment.[3] The test statistic is an *F* statistic, which compares treatment variance to cluster variance. The *F* statistic is defined as:

$$Fstatistic = \frac{(MS_{treatment})}{(MS_{cluster})} \qquad [7.6]$$

Note that as the number of clusters J increases without bound, the *F* statistic converges to the ratio of expected mean squares, which is defined as:

[3] This is the same result we would obtain using a two-level hierarchical linear model (Equations 1and 2) estimated by means of restricted maximum likelihood.

$$\frac{E(MS_{treatment})}{E(MS_{cluster})} = \frac{n\tau + \sigma^2 + nJ\gamma_{01}^2/4}{n\tau + \sigma^2} = 1 + \frac{nJ\gamma_{01}^2/4}{n\tau + \sigma^2} \quad [7.7]$$

and can be rewritten as:

$$\frac{E(MS_{treatment})}{E(MS_{cluster})} = 1 + \lambda \text{ where } \lambda = \frac{nJ\gamma_{01}^2/4}{n\tau + \sigma^2}. \quad [7.8]$$

If the null hypothesis is true, the *F* statistic follows a central *F* distribution with 1 degree of freedom for the numerator and *J*-2 degrees of freedom for the denominator. Under the central *F* distribution, we would expect the *F* statistic to be approximately 1. In other words, there is no variation between treatments so $\gamma_{01} \approx 0$ and the $nJ\gamma_{01}^2/4$ term in the numerator of the expected mean square ratio goes towards 0. We see that if $\lambda = 0$ the ratio of expected mean squares thus reduces to $\frac{E(MS_{treatment})}{E(MS_{cluster})} = \frac{n\tau + \sigma^2}{n\tau + \sigma^2} = 1 + \lambda = 1.$

If the null hypothesis is false so that there is a treatment difference, that is $\gamma_{01} \neq 0$, the *F* statistic follows a non-central *F* distribution with 1 degree of freedom for the numerator and *J*-2 degrees of freedom for the denominator. Then the ratio of expected mean squares becomes the non-central *F* distribution, characterized by a non-centrality parameter, λ (See Equation 7.8). λ can be rewritten as:

$$\lambda = \frac{\gamma_{01}^2}{4(\tau + \sigma^2/n)/J} \quad [7.9]$$

Note that λ, known as the non-centrality parameter, is the ratio of the squared main effect to the variance of the estimate of the treatment effect. Equation 9 clearly shows that the non-centrality parameter, λ, is a function of γ_{01}, *n*, *J*, τ, and σ^2.

The non-centrality parameter is strongly related to the power of the test. As λ increases, the power increases. Increasing the treatment effect increases λ. Thus, if we are trying to detect a larger difference in means, λ increases and so the power also increases. Note that the denominator is identical to the variance of the treatment effect (Equation 7.5). So to increase λ we could decrease the variance of the main effect of treatment. Because the standard error of the treatment effect is more commonly discussed, instead of referring to the variance of the main effect of treatment, we often refer to the standard error of the main effect of treatment, which is simply:

$$SE(\hat{\gamma}_{01}) = \sqrt{\frac{4(\tau + \sigma^2 / n)}{J}} \qquad [7.10]$$

From equation 7.10, we can see that the sample sizes affect the standard error and hence the power of the test. In general, increasing *n* decreases the standard error of the treatment effect thus increasing the power. However, at some point, increasing *n* without increasing the number of clusters, *J*, provides no further benefit. Thus as $n \to \infty$, we can see that for Equation 10, $SE(\hat{\gamma}_{01}) = 2\sqrt{\tau / J}$, will not be zero unless $\tau = 0.$ Also, as the total number of clusters, *J*, increases, the power to detect significant differences also increases. As *J* increases towards infinity, the power approaches 1 regardless of *n*. This is because as *J* increases towards infinity, the standard error (7.10) gets infinitely small. This causes the non-centrality parameter to increase towards infinity, which results in the power approaching 1. Intuitively this makes us think that we should just continue to increase *J* until the desired power is achieved. However, increasing *J* or adding additional clusters may not be feasible due to budgetary constraints.

7.3 Standardized notation

We standardize the notation to give a more meaningful definition for the parameters in the model and to facilitate the power analysis. First, we redefine the variability in terms of the intra-class correlation coefficient, ρ. The intra-class correlation, ρ, is a ratio of the variability between clusters to the total variability:

$$\rho = \frac{\tau}{\tau + \sigma^2} \qquad [7.11]$$

where τ is the variation between clusters;

σ^2 is the variation within clusters; and

$\tau + \sigma^2$ is the total variation.

For US data sets on school achievement, ρ typically ranges between 0.15 and 0.25 (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Schochet, 2008). In school-based interventions design to improve mental health, ρ will generally be smaller, in the range of 0.01 to 0.05 (Murray & Short, 1995) . Because $\tau + \sigma^2$ is the total variation, we can constrain it to be 1. Algebraic manipulation of the formula then reveals $\rho = \tau$ and $1 - \rho = \sigma^2$. As ρ increases we know more of the variation is due to between-cluster

variability. Replacing τ and σ^2 with ρ and 1- ρ in the standard error formula (equation 7.10), the standard error of the main effect of treatment can be rewritten as:

$$SE(\hat{\gamma}_{01}/\sqrt{(\tau+\sigma^2)} = \sqrt{\frac{4(\tau+\sigma^2/n)/(\tau+\sigma^2)}{J}} = \sqrt{\frac{4(\rho+(1-\rho)/n)}{J}}. \qquad [7.12]$$

From equation 7.12, we can see that increased values of ρ increase the standard error thus decreasing the power. Also, as ρ increases, the effect of *n* decreases. Therefore, if there is a lot of variability between clusters, we gain more power by increasing the number of clusters sampled. The key idea for ρ is that power increases as ρ decreases for a fixed *n* and *J*.

Next, we can standardize the true treatment effect. Because data for different experiments is collected in different scales, standardizing the data makes the results meaningful to any researcher, not just someone who is familiar with a particular data set. We define the standardize effect size, δ, as the population means difference of the two groups divided by the standard deviation of the outcome:

$$\delta = \frac{\gamma_{01}}{\sqrt{\tau + \sigma^2}} \qquad [7.13]$$

Where $\gamma_{01} = \mu_E - \mu_C$;

μ_E is the population mean for the experimental group; and

μ_C is the population mean for the control group;

Given σ^2 and τ, the standardized effect size, δ, is estimated by:

$$\hat{\delta} = \frac{\bar{y}_E - \bar{y}_C}{\sqrt{\tau + \sigma^2}} \qquad [7.14]$$

The researcher must specify a desired minimum effect size to calculate the power of the test. Quantifying the treatment effect is not easy. It depends on the context of the study, the sample, and the outcomes in the study (Bloom, Hill, Black, & Lipsey, 2007).

Recall that the power of the test is driven by the non-centrality parameter, λ (equation 7.9). We can redefine λ in standardized notation as shown below:

$$\lambda = \frac{\gamma_{01}^2/(\tau+\sigma^2)}{4(\tau+\sigma^2/n)/J(\tau+\sigma^2)} = \frac{nJ\delta^2/4}{n\rho+(1-\rho)} = \frac{J\delta^2}{4(\rho+(1-\rho)/n)} \qquad [7.15]$$

This allows us to calculate the power of the as a function of *n*, *J*, δ, and ρ.

7.4 Using a covariate to increase power

From a power perspective, including a covariate can be extremely helpful because if the covariate is strongly correlated with the outcome, it can greatly increase the precision of the estimate and hence the power of the study. We focus specifically on including a covariate at the cluster level. This may be an aggregated covariate, such as pre-test scores aggregated across schools or school SES. Empirical work has shown that similar gains in power for including an individual level or cluster level covariate (Bloom, Richburg-Hayes, & Black, 2007). Because it is generally less time consuming and less expensive to collect a cluster level covariate, we focus on level-2 covariates.

When we include a covariate in the design, there is an additional component that influences the power of the test: the strength of the correlation between the covariate and the true cluster mean outcome. The strength of the correlation between the covariate and the true cluster mean is denoted $\rho_{x\beta_0}$. We adopt this notation because β_{0j} is the true mean outcome for cluster *j*, and X_j is the covariate. The residual level-2 variance, or unexplained variance after accounting for the covariate, is denoted $\tau_{|x}$. As we will see later, the stronger the correlation, $\rho_{x\beta_0}$, the smaller the conditional level 2 variance, $\tau_{|x}$, compared to the unconditional level 2 variance, τ, and the greater the benefit of the covariate in increasing precision and power. Let's take a closer look at the model with a cluster-level covariate.

7.5 The model with a cluster-level covariate

In hierarchical form, the level-1 model for a cluster randomized trial with a cluster-level covariate is the same as the model in equation 1. The level-2 model, or cluster-level model differs from a simple cluster randomized trial because it includes a term for the cluster-level covariate. The model with the covariate is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + \gamma_{02}X_j + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{|x}) \qquad [7.16]$$

where γ_{00} is the grand mean;

γ_{01} is the mean difference between the treatment and control group or the main effect of treatment;

γ_{02} is the regression coefficient for the cluster-level covariate;

W_j is the treatment contrast indicator, ½ for treatment and -½ for control;

X_{ij} is the cluster-level covariate, centered around its group mean;

u_{0j} is the random effect associated with each cluster; and

$\tau_{|x}$ is the residual variance between clusters.

Note that the between-cluster variance, $\tau_{|x}$, is now the residual variance conditional on the cluster level covariate X. For simplicity, we assume there is no interaction between the cluster level covariate, X, and the treatment group, W. This is an assumption that can be relaxed and in general should be checked given that a researcher is interested in how the treatment effect may vary at different levels of the covariate.

7.6 Testing the treatment effect (including a cluster-level covariate)

Similar to the cluster randomized trial without a covariate, we are interested in the main effect of treatment, or the difference between the treatment average and control average adjusting for the covariate. However, now it is estimated by:

$$\hat{\gamma}_{01} = \bar{Y}_E - \bar{Y}_C - \hat{\gamma}_{02}(\bar{X}_E - \bar{X}_C) \qquad [7.17]$$

where $\bar{Y}_E$ is the mean for the experimental group;

$\bar{Y}_C$ is the mean for the control group;

$\bar{X}_E$ is the covariate mean for the experimental group; and

$\bar{X}_C$ is the covariate mean for the control group.

Note that the estimated main effect of treatment looks like the estimated effect without the covariate except that here we are adjusting for treatment group differences in the covariate. For normally distributed covariates, the variance of the main effect is (Raudenbush, 1997):

$$Var(\hat{\gamma}_{01}) = \frac{4(\tau_{|x} + \sigma^2 / n)}{J}\left[1 + \frac{1}{J-4}\right] \qquad [7.18]$$

where n is the total number of subjects;

J is the total number of clusters; and

$\tau_{|x}$ is the conditional level 2 variance, $(1-\rho^2_{x\beta_0})\tau$.

If the data are balanced, we can use the results of a nested analysis of covariance with random effects for clusters and fixed effects for the treatment and covariate. The test statistic is

an *F* statistic, which compares adjusted treatment variance to the adjusted cluster variance. The *F* statistic is defined as:

$$Fstatistic = \frac{MS_{treatment}}{MS_{clusters}},$$

where $MS_{treatment}$ and $MS_{cluster}$ are now adjusted for the covariate.

Note that the *F* statistic converges to the ratio of expected mean squares, defined as:

$$\frac{E(MS_{treatment})}{E(MS_{clusters})} = 1 + \lambda_x$$

The *F* test follows a non-central *F* distribution, $F(1,J\text{-}3, \lambda_x)$ in the case of a cluster-level covariate where the non-centrality parameter, λ_x, is:

$$\lambda_x = \frac{J\gamma_{01}^2}{4(\tau_{|x} + \sigma^2/n)} \qquad [7.19]$$

and

$$\tau_{|x} = (1 - \rho_{x\beta_o}^2)\tau \ .$$

From equations 7.18 and 7.19, we can see that the stronger the correlation, $\rho_{x\beta_0}$, the smaller $\tau_{|x}$, and the greater the increase in the power of the test.

The non-centrality parameter with and without the covariate are closely related. If the correlation between the covariate and the cluster level mean is 0, $\tau_{|x}$ reduces to τ and the non-centrality parameter reduces to λ, the non-centrality parameter in the case of no covariate. Although we are reducing the between cluster variance, one consequence of including a covariate is that we lose one degree of freedom. In the case of no covariate, the *F* test follows a non-central *F* distribution, $F(1, J\text{-}2, \lambda)$ whereas in the covariate case we have $F(1, J-3; \lambda_x)$. This may be a potential problem in a study with a small number of clusters.

The non-centrality parameter can be defined in standardized notation. Recall that in equation 7.19 we define the non-centrality parameter as $\lambda_x = \frac{J\gamma_{01}^2}{4(\tau_{|x} + \sigma^2/n)}$. Defining $\delta = \frac{\gamma_{10}}{\sqrt{\tau + \sigma^2}}$ and $\rho = \frac{\tau}{\tau + \sigma^2}$ as we did in the unconditional case, but also including R^2, the

estimated percent of variance explained by the cluster-level covariate, e can rewrite λ_x as a function of δ, ρ and $R^2_{Level\,2}$ as shown below:

$$\lambda_x = \frac{J\gamma_{01}^2}{4(\tau_{|x} + \sigma^2 / n)} = \frac{J\delta^2}{4[(1-R^2_{Level\,2})\rho + (1-\rho)/n]} \qquad [7.20]$$

The only difference in the non-centrality parameter in the case of the cluster level covariate is the correction factor, $(1-R^2_{Level\,2})$. The correction factor only affects τ, the between-cluster variation since the covariate is a cluster-level covariate. As the correlation between the covariate and the cluster level means increases, the unconditional intra-class correlation decreases. This results in an increase in the value of the non-centrality parameter and therefore an increase in the power of the test.

7.8 Using the Optimal Design for two-level cluster randomized trials

The menu for the 2-level CRT is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 2. In this chapter we focus on continuous outcomes. The available options are shown below.

Power on y-axis (continuous outcomes)

- Power vs. cluster size (*n*)
- Power vs. number of clusters (*J*)
- Power vs. intra-class cluster correlation (rho)
- Power vs. effect size (delta)
- Power vs. proportion of explained variation by level 2 covariate (R2)[4]

MDES on y-axis (continuous outcomes)

- MDES vs. cluster size (*n*)
- MDES vs. number of clusters (*J*)
- MDES vs. intra-class cluster correlation (rho)
- MDES vs. power (*P*)
- MDES vs. proportion of explained variation by level 2 covariate (R2)

[4] Note that this differs from Version 1.0 of the program. In Version 1.0, the program asked for a covariate correlation. In Version 2.0, the program asks for the proportion of explained variation by the level 2 covariate, R^2.

The first set of options present the power on the y-axis and either the cluster size, number of clusters, intraclass correlation, effect size, or proportion of explained variation by level-2 covariate to vary on the x-axis. The second set of options present the MDES on the y-axis and either the cluster size, number of clusters, intraclass correlation, power, or proportion of explained variation by level-2 covariate to vary on the x-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

7.9 Example

Suppose a team of researchers develop a new literacy program for 3rd graders. The founders of the new program propose that students who participate in the program will have increased reading achievement. They plan to test students who participate in the new program (experimental group) and students who participate in the regular program (control group) using a standardized reading test to determine if students using the new program score higher. The researchers have access to last years 3rd grade average reading test scores for each school. Past data reveals that last years scores explain 49% of the variation in test scores. The researchers want to design a cluster randomized trial with students nested within schools where schools are the unit of randomization. Section 7.10 presents a scenario in which the power determination approach for conducting a power analysis is most applicable and the details of how to do the power analysis using OD. Section 7.11 presents a scenario in which the effect size approach is most applicable the details of how to do the power analysis using OD.

7.10 Power determination approach for conducting a power analysis

Based on past studies, the researchers expect about 20 percent of the variation to lie between schools and are interested in detecting an effect size of at least 0.25 with adequate power. Assuming that 20 students are willing to participate in the study from each school, how many schools (clusters) are necessary to achieve power = 0.80? How many clusters are required after including the cluster-level covariate which explains 49% of the variation in test scores?

In Scenario 1, the number of clusters, *J*, is unknown. As a result, we want to select the power vs. number of clusters (*J*) option. This allows the number of clusters to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 2→ Power on y-axis (continuous outcome) → Power vs. number of clusters (*J*) as shown in Figure 7.1.

The blank screen for Power vs. number of clusters (*J*) is in Figure 7.1.

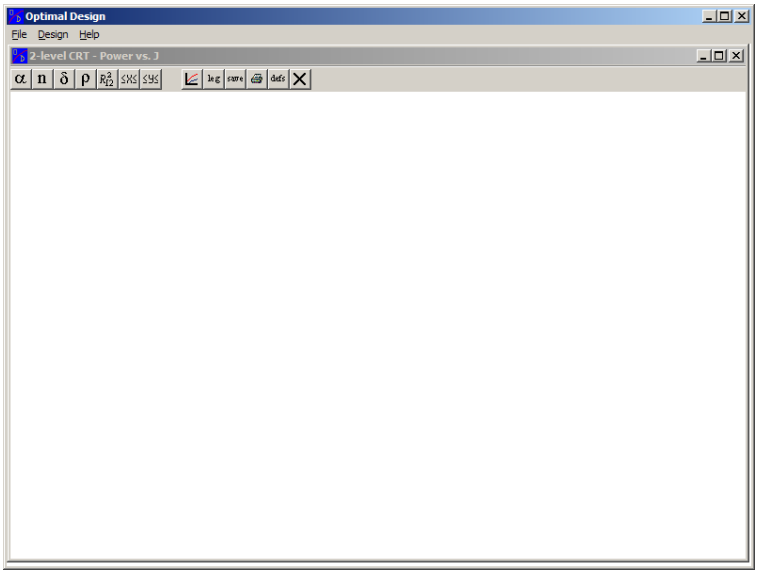


Figure 7.1. Initial screen.

The toolbar at the top includes the parameters required for calculating the power, sample size within cluster (n), effect size (δ), intraclass correlation (ρ), and explained proportion of variance by covariate. The number of clusters (*J*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 20 (the default value). By clicking on n(1) = 20, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on δ. Set delta (1) = 0.25 and delete delta (2).

Step 4: Click on ρ. Set rho (1) = 0.20 and delete rho (2). The resulting power curve appears in figure 7.2.

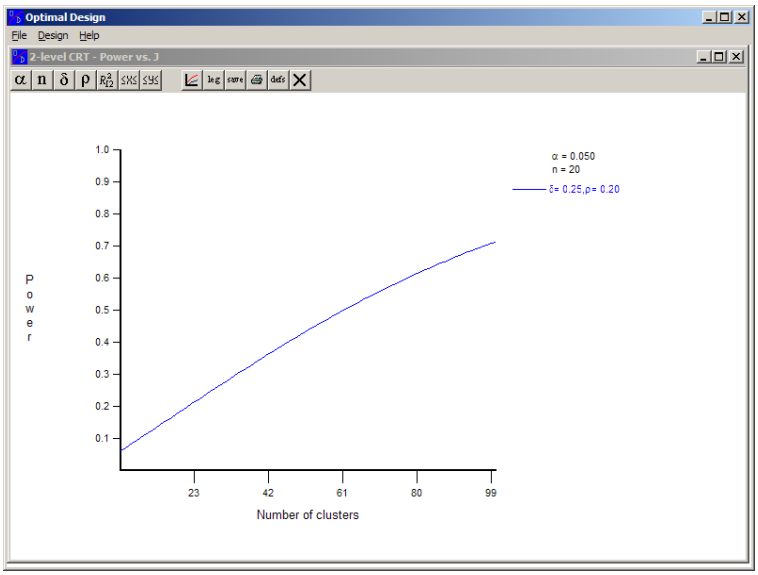


Figure 7.2. Resulting power curve.

We can see that the curve does not extend to power of 0.80 so we need to extend the x-axis.

Step 5: Click on <x<. Set the maximum to 150. The extended power curve appears in Figure 7.3.

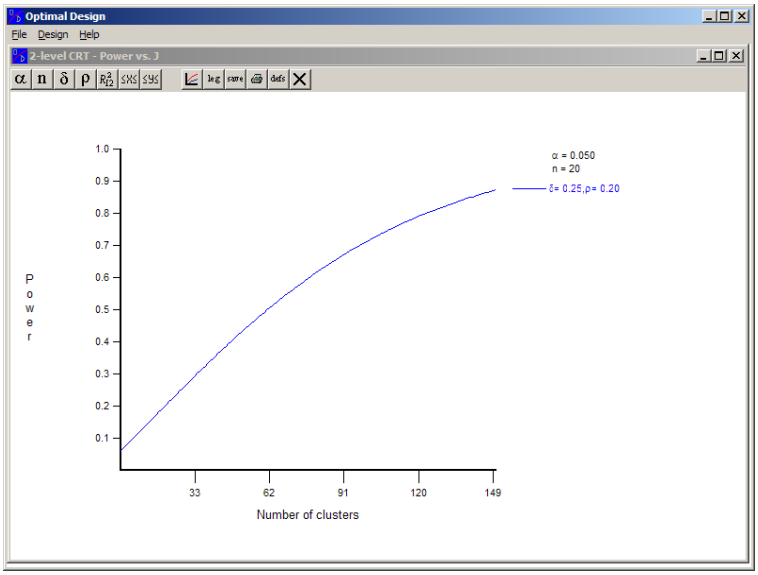


Figure 7.3. Extended power curve.

Clicking along the power curve, we can see that 122 total clusters are required for the study, 61 in the treatment condition and 61 in the control condition. However, we have not accounted for the cluster-level covariate yet, which is a strategy for increasing the precision of the estimate and the power.

Step 6: Click on R. Set R2 – 2 equal to 0.49. Two power curves appear as shown in Figure 7.4.

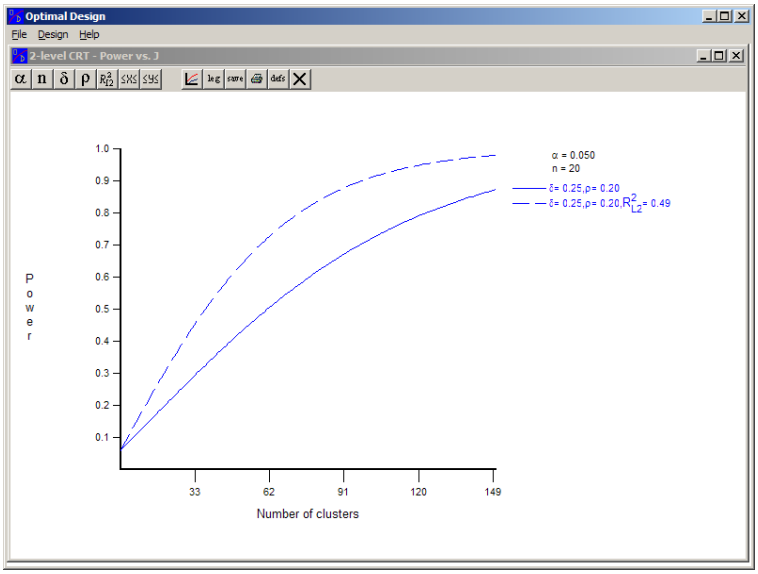


Figure 7.4. Two power curves.

According to the key, the dotted trajectory is the power curve when the covariate is included. In this case, approximately 74 total clusters are needed, a reduction of 48 clusters, which may greatly reduce the cost of the study.

The example provided in this section placed the sample size on the x-axis. However, any of the other parameters could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

7.11 Effect size approach for conducting a power analysis

The researchers conducting the study are limited to 60 schools, 30 in the treatment and 30 in the control group with 20 students per school. Based on past studies, they expect about 20 percent of the variation to lie between schools. What is the MDES the researchers can find with power = 0.80? Assuming the cluster-level covariate explains 49% of the variation in test scores, what is the MDES?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. total number of clusters (*J*). This will allow the user to see how the MDES changes as a function of the total number of clusters holding all other parameters constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 2 →Power on y-axis (continuous outcome) → MDES vs. total number of clusters (*J*) as shown in Figure 7.5.

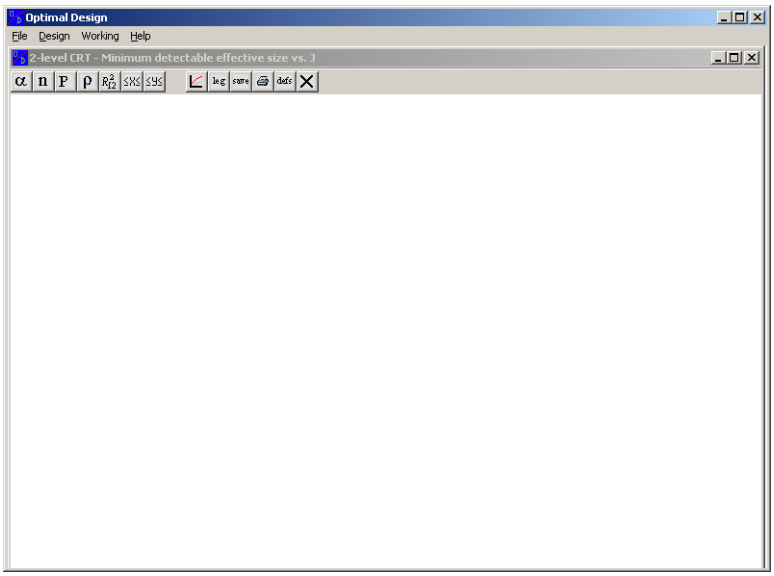


Figure 7.5. Blank screen for MDES vs. total number of clusters.

Step 2: Click on n. Set n(1) = 20 (the default value). By clicking on n(1) = 20, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *P*. Set *J* (1) = 0.80.

Step 4: Click on ρ. Set rho (1) = 0.20 and delete rho (2). The resulting power curve appears in Figure 7.6.

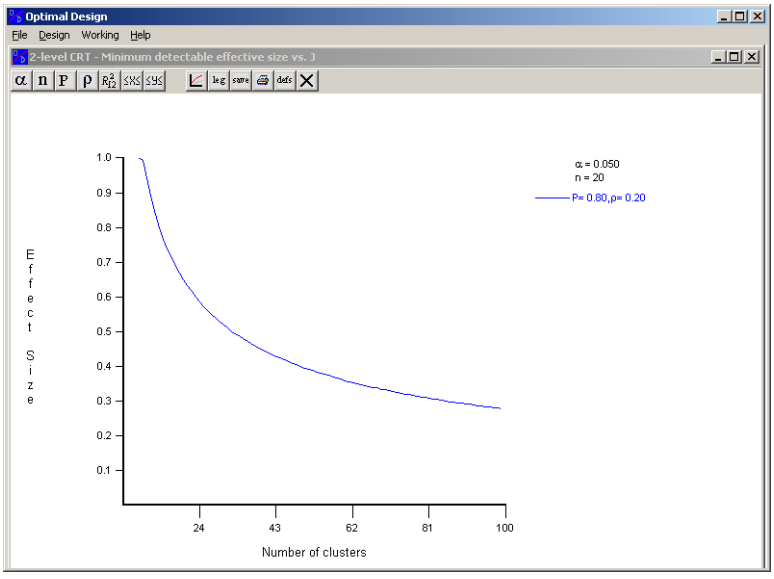


Figure 7.6. Power curve with specified parameters.

Clicking along the trajectory we can see that for $J = 60$, the MDES is approximately 0.36. Let's see what happens when we add a cluster-level covariate.

Step 5: Click on R. Set R – 2 = 0.49. Figure 7.7 displays the two power curves.

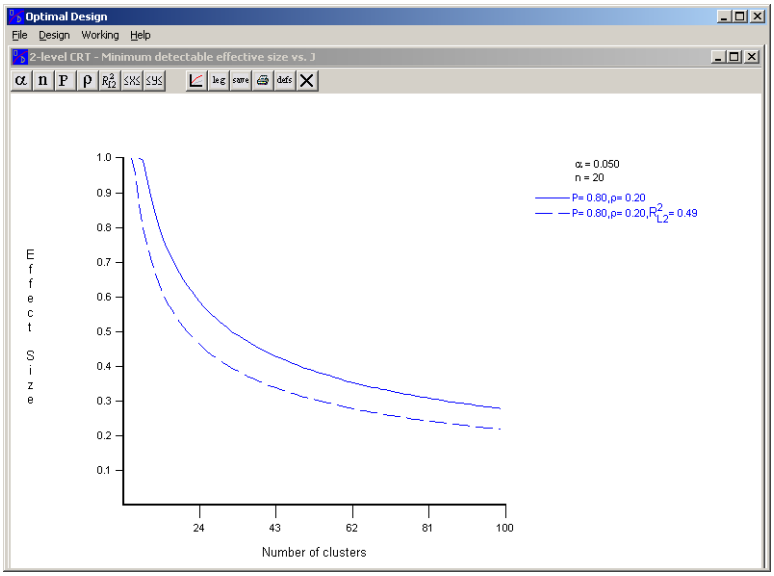


Figure 7.7. Power curve with cluster-level covariate.

Clicking along the dotted trajectory, the MDES with $J = 60$ is 0.28, reducing the MDES by 0.08 effect size units.

The examples in this section are meant to provide at guide to users for how to use the 2-level CRT. We described Power vs. number of clusters (*J*) and MDES vs. number of clusters (*J*) The other options function similarly, and simply place a different parameter on the x-axis. The choice of which module is most appropriate depends on the unknown parameters. However, all modules yield the same results if identical parameters are used so the choice depends on what module is most closely aligned with the known and unknown parameters in a study.

8.0 Three-level cluster randomized trials (3-level CRT)

Three-level cluster randomized trials are studies in which individuals are nested within clusters, clusters are nested within sites, and sites are randomly assigned to the treatment or control condition. For example, students are nested within classrooms and classrooms are nested within schools. Suppose a team of researchers are interested in the effectiveness of a comprehensive school reform (CSR) on math outcomes. They decide to randomly assign schools to either the new CSR or the current program. They plan to test students from multiple classrooms within each school. In this case, schools are the unit of randomization, but the students are nested within classrooms which are nested within schools making this a 3-level CRT. The importance of including the classroom level is that it allows us to examine the variability between classrooms. If we suspect that there will be significant differences among classrooms, or teachers, it is important to include this level in the design and analysis. The additional level makes the power for a 3-level CRT more complex than for a 2-level CRT. We begin by examining the underlying statistical models.

8.1 The model

We can represent the data from this design as persons nested within clusters nested within sites. The level 1, or person-level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \qquad [8.1]$$

where $i = 1,...,n$ persons per cluster

j=1,…,J clusters per site

k=1,…,K sites

π_{0jk} is the mean for cluster j in site k

e_{ijk} is the error associated with each person

σ^2 is the within-cluster variance.

The level-2 model, or cluster-level model, is:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi) \qquad [8.2]$$

where β_{00k} is the mean for site k

r_{0jk} is the random effect associated with each cluster

τ_π is the variance between clusters within sites.

The level-3 model, or site-level model, is:

$$\beta_{00k} = \gamma_{000} + \gamma_{001} W_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}}) \qquad [8.3]$$

where γ_{000} is the estimated grand mean

γ_{001} is the treatment effect ("main effect of treatment")

W_k is 0.5 for treatment and –0.5 for control

u_{00k} is the random effect associated with each site mean

$\tau_{\beta_{00}}$ is the residual variance between site means.

Note that the randomization in this design occurs at level 3.

8.2 Testing the treatment effect

The treatment effect is estimated at level 3 and is denoted γ_{001}. Given a balanced design, it is estimated by:

$$\hat{\gamma}_{001} = \bar{Y}_E - \bar{Y}_C \qquad [8.4]$$

where $\bar{Y}_E$ is the mean for the experimental group

$\bar{Y}_C$ is the mean for the control group.

Because of the nested structure of the data, we sum over clusters and sites in order to estimate the treatment effect. The variance of the estimated treatment effect combines the variance at all three levels, the variance between-site means, $\tau_{\beta_{00}}$, the within-site or between-cluster variance, τ_π, and the within-cluster or between-person variance, σ^2. The variance of the treatment effect is estimated by:

$$Var(\hat{\gamma}_{001}) = \frac{4[\tau_{\beta_{00}} + (\tau_\pi + \sigma^2 / n) / J]}{K} \qquad [8.5]$$

If the data are balanced, we can use the results of a nested analysis of variance with random effects for the clusters and sites and a fixed effect for the treatment. The test statistic is an *F* statistic. The *F* test follows a non-central *F* distribution, $F(1, K\text{-}2; \lambda)$. Below is the noncentrality parameter for the test, λ, which is the ratio of the squared-treatment effect to the variance of the treatment effect estimate.

$$\lambda = \frac{\gamma_{001}^2}{\hat{Var}(\gamma_{001})} = \frac{K\gamma_{001}^2}{4[\tau_{\beta_{00}} + (\tau_\pi + \sigma^2/n)/J]} \quad [8.6]$$

Increasing the noncentrality parameter increases the power to detect the treatment effect. Let's examine how the researcher can increase the noncentrality parameter to increase the power of the test. Because this model assumes no covariates, we cannot reduce any of the variance components so $\tau_{\beta_{00}}$, τ_π, and σ^2 are not under the control of the researcher. The only remaining components of the noncentrality parameter are the sample size and the size of the treatment effect. The size of the treatment effect is often based on theory, past studies, or a pilot study which means the researcher cannot inflate the size of the treatment effect to increase power without decreasing the theoretical or practical conclusions of the study. Thus increasing the sample size is the primary option for increasing the power. From equation 8.6, we can see that increasing the number of sites, *K*, is the most effective strategy to increase the power, followed by the number of clusters, *J*, and finally the number of persons per cluster, *n*.

8.3 Standardized notation

Thus far we have focused on the unstandardized model. However researchers typically discuss standardized effect sizes. In the standardized model, without loss of generality, we set the sum of the within-cluster variance, σ^2, the between-cluster variance, τ_π, and the between-site variance for the site means, $\tau_{\beta_{00}}$, equal to 1. Since we use three components of variance to standardize the model, we have two intra-class correlations, ρ_{level2} and ρ_{level3}. The first intra-class correlation, ρ_{level2}, corresponds to the between-cluster variance relative to the total variance, $\rho_{level2} = \frac{\tau_\pi}{\tau_\pi + \tau_{\beta_{00}} + \sigma^2}$. The second intra-class correlation, ρ_{level3}, is the between-site variance relative to the total between and within site variance, $\rho_{level3} = \frac{\tau_{\beta_{00}}}{\tau_{\beta_{00}} + \tau_\pi + \sigma^2}$.

In standardized notation, the non-centrality parameter, λ, can be rewritten as:

$$\lambda = \frac{K\delta^2}{4\{\rho_{level3} + [\rho_{level2} + (1 - \rho_{level2} - \rho_{level3})/n]/J\}} \quad [8.7]$$

where δ is the standardized main effect of treatment, $\delta = \frac{\gamma_{001}}{\sqrt{\tau_{\beta_{00}} + \tau_\pi + \sigma^2}}$.

Since the total variance is constrained to 1, users of the Optimal Design can think of ρ_π and ρ_β as estimates of the proportion of variance at level 2 and level 3.

8.4 Using a covariate to increase power

Often a covariate may be available to the researcher. The researchers can use this information to reduce the level-3 variability, or the between-site variance and increase the power of the test. For the 3-level CRT, we restrict our attention to level-3 covariates. Often they are less expensive to collect, more readily available, and can increase the precision by a similar margin of a lower level covariate. Thus if a lower level covariate is available, we assume it is aggregated to level-3. We also assume it has met the assumptions for inclusion. Including a site-level covariate will not effect the between-cluster variability, τ_π or the within-cluster variability, σ^2. We use *S* to denote a site-level covariate in the model. The proportion of variance explained by the site-level covariate is defined as $\rho^2_{s\beta_{00}}$. The remaining sections in this chapter revisit the model, treatment effect, and standardized notation assuming the availability of a site-level covariate.

8.5 The model with a covariate

Levels 1 and 2 of the model with a site-level covariate are identical to the level 1 and 2 equations (equations 1 and 2) for the case with no covariate. This is because inclusion of a site-level covariate does not effect the variability in the lower levels in the model. The new level 3, or site level model is:

$$\beta_{00k} = \gamma_{000} + \gamma_{001}W_k + \gamma_{002}S_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}|s}) \qquad [8.8]$$

$$\text{Note: } \tau_{\beta_{00k}|s} = (1 - \rho^2_{s\beta_{00}})\tau_{\beta_{00}}$$

where γ_{000} is the estimated grand mean

γ_{001} is the treatment effect ("main effect of treatment")

γ_{002} is the regression coefficient for the level 3 covariate

W_k is 0.5 for treatment and –0.5 for control

S_k is the level 3 covariate

u_{00k} is the random effect associated with each site mean

$\tau_{\beta_{00}|s}$ is the residual variance between site means conditional on the site-level covariate

Note that the level 3 variance is adjusted for the covariate. The smaller variance will increase the precision of the estimate thus increasing the power of the test.

Given a balanced design, the main effect of treatment is estimated as the difference in the treatment and control groups adjusted for the site-level covariate:

$$\hat{\gamma}_{001} = \bar{Y}_E - \bar{Y}_C - \hat{\gamma}_{002}(\bar{S}_E - \bar{S}_C). \quad [8.9]$$

The variance of the treatment effect is:

$$Var(\hat{\gamma}_{001} \mid S) = \frac{4[\tau_{\beta_{00}|s} + (\tau_\pi + \sigma^2 / n)/J]}{K}. \quad [8.10]$$

Note that only the between-site variance, $\tau_{\beta_{00}}$, is adjusted for inclusion of the covariate since it is at the covariate is at the site level.

Similar to the case with no covariate, to test the main effect of treatment we use an F-statistic which follows a non-central F distribution, F(1, *K*-3, λ_s) where:

$$\lambda_s = \frac{K\gamma_{001}^2}{4[\tau_{\beta_{00}|s} + (\tau_\pi + \sigma^2 / n)/J]}. \quad [8.11]$$

The noncentrality parameter for the test for the main effect of treatment looks similar to equation 6, the case with no covariate, except that the level 3 variance and the estimate of the treatment effect are adjusted for the cluster level covariate. Note that reducing the variability at level 3 gives the researcher another tool for increasing the noncentrality parameter and increasing the power. In cases when the between-site variance accounts for a high proportion of the variance, finding a site-level covariate that is highly correlated with the site-level outcome can be very beneficial. It may also help reduce the number of sites necessary to achieve a specified power, which can reduce the cost of the study.

Following the same logic as the three level model with no covariates, it is important to standardize the model. The noncentrality parameter expressed in standardized notation is:

$$\lambda_s = \frac{K\delta^{*2}}{4\{\rho^*_{level3} + [\rho^*_{level2} + (1 - \rho^*_{level2} - \rho^*_{level3})/n]/J\}} \quad [8.12]$$

where

ρ^*_{level2} is the intra-class correlation $\rho^*_{level2} = \frac{\tau_\pi}{\tau_\pi + \tau_{\beta_{00}|s} + \sigma^2}$, or the proportion of variance among clusters relative to the total variation conditional on the level-3 covariate.

ρ^*_{level3} is the intra-class correlation $\rho^*_{level3} = \dfrac{\tau_{\beta_{00}|s}}{\tau_{\beta_{00}|s} + \tau_\pi + \sigma^2}$, or the proportion of variance among sites relative to the total variation conditional on the level-3 covariate.

δ^* is the standardized main effect of treatment conditional on the level-3 covariate, $\delta^* = \dfrac{\delta}{\sqrt{\tau_{\beta_{00}|s} + \tau_\pi + \sigma^2}}$.

Because the conditional standardized quantities, ρ^*_{level2}, ρ^*_{level3}, and δ^*, are frequently unknown, the program asks the user to enter the unconditional parameters. The program calculates the conditional standardized values based on the value the user specifies for the percent of variance reduction at level 3, R^2_{level3}.

8.6 Using the Optimal Design for three-level cluster randomized trials

This section focuses on how to use the Optimal Design software to design a three-level cluster randomized trial with a continuous outcome. Section 8.7 presents an example. The remaining two sections explore how to conduct a power analysis using 1) the power determination approach, and 2) the effect size approach.

The menu for the 3-level CRT is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 3. In this chapter we focus on continuous outcomes thus the first two options are shown below.

Power on y-axis (continuous outcomes)

Power vs. cluster size (*n*)

Power vs. number of clusters per site (*J*)

Power vs. number of sites (*K*)

Power vs. effect size (delta)

Power vs. proportion of explained variation by level-three covariate (R2)

MDES on y-axis (continuous outcomes)

MDES vs. cluster size (*n*)

MDES vs. number of clusters per site (*J*)

MDES vs. number of sites (*K*)

MDES vs. power (*P*)

MDES vs. proportion of explained variation by level-three covariate (R2)

The first set of options present the power on the y-axis and either the cluster size, number of clusters per site, number of sites, effect size, or proportion of explained variation by level-3 covariate to vary on the x-axis. The second set of options present the MDES on the y-axis and either the cluster size, number of clusters per site, number of sites, power, or proportion of explained variation by level-3 covariate to vary on the x-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

8.7 Example

Suppose a team of researchers are interested in testing the effectiveness of a whole school reform model. They plan to implement the program at the school level and set up an experiment which randomly assigns schools to either the new reform model or the current practice. They suspect there may be teacher level differences so they plan to set up a three level study with students nested within teachers nested within schools. The primary outcome of interest is math achievement. The researchers plan to test students who participate in the new program (experimental group) and students who participate in the regular program (control group) using a standardized reading test to determine if students using the new program score higher. The researchers have access to last years 1st grade average math test scores for each school. Past data reveals that last years scores explain 49% of the variation in test scores. Section 8.8 presents a scenario in which the power determination approach for conducting a power analysis is most applicable and the details of how to do the power analysis using OD. Section 8.9 presents a scenario in which the effect size approach is most applicable the details of how to do the power analysis using OD.

8.8 Power determination approach for conducting a power analysis

Based on past studies, the researchers expect about 13 percent of the variation to lie between schools and 7 percent of the variation to lie between classrooms within schools. They are interested in detecting an effect size of at least 0.25 with adequate power. Assuming that 20 students are willing to participate in the study from each classroom, and there are 12 teachers per school (assume an elementary school where each teacher teaches one class) how many schools

(sites) are necessary to achieve power = 0.80? How many schools are required after including the school-level covariate which explains 49% of the variation in test scores?

In Scenario 1, the number of sites, *K*, is unknown. As a result, we want to select the power vs. number of sites (*K*) option. This allows the number of clusters to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 3 → Power on y-axis (continuous outcome) → Power vs. number of sites (*K*) as shown in Figure 8.1.

The blank screen for Power vs. number of sites (*K*) is in Figure 8.1.

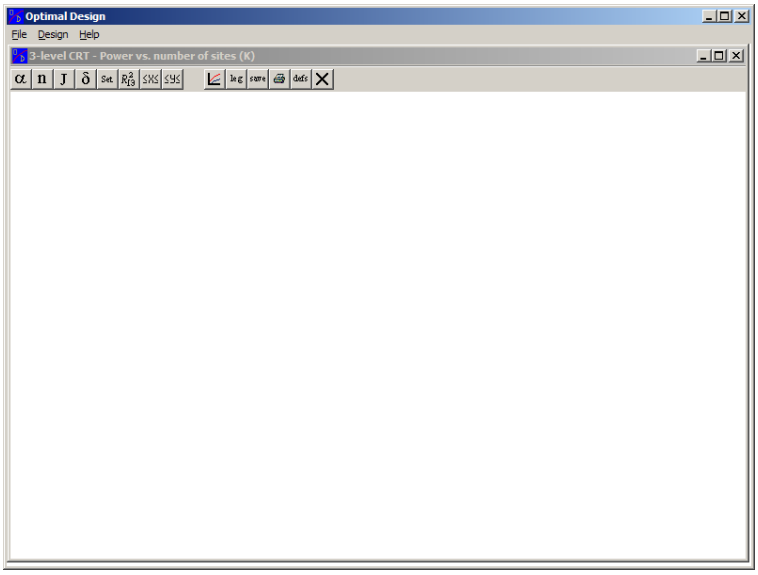


Figure 8.1. Initial screen for cluster randomized trials.

The toolbar at the top includes the parameters required for calculating the power: sample size within cluster (n), number of clusters per site (*J*), effect size (δ), and explained proportion of variance by covariate. The set button asks the user to specify the level-2 intraclass correlation, ρ_π, and the level-3 intraclass correlation, ρ_β. The number of sites (*K*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 20. By clicking on n(1) = 20, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J*(1) = 12. This is the number of classrooms per school.

Step 4: Click on δ. Set delta (1) = 0.25 and delete delta (2).

Step 5: Click on set. Set ρ_π = 0.07 and ρ_β = 0.13. The resulting power curve appears in figure 8.2.

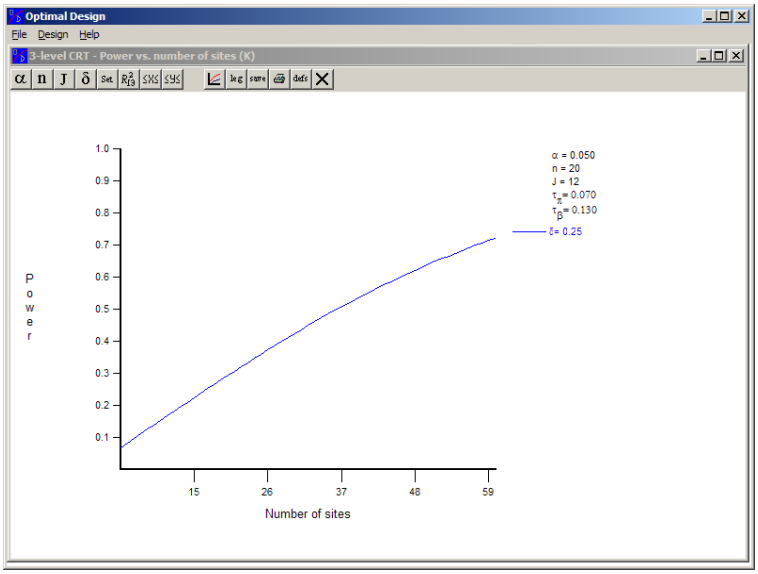


Figure 8.2. Power curve with specified parameters.

We can see that the curve does not extend to power of 0.80 so we need to extend the x-axis.

Step 6: Click on <x<. Set the maximum to 100. The extended power curve appears in Figure 8.3.

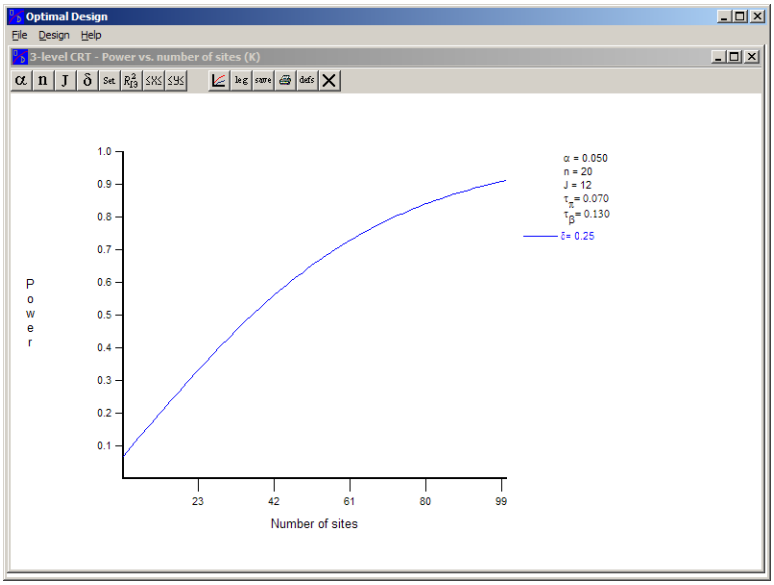


Figure 8.3. Extended power curve.

Clicking along the power curve, we can see that 72 total schools are required for the study, 36 in the treatment condition and 36 in the control condition. However, we have not accounted for the cluster-level covariate yet, which is a strategy for increasing the precision of the estimate and the power of the study.

Step 7: Click on R. Set R^2 – (2) = to 0.49. Two power curves appear as shown in Figure 8.4.

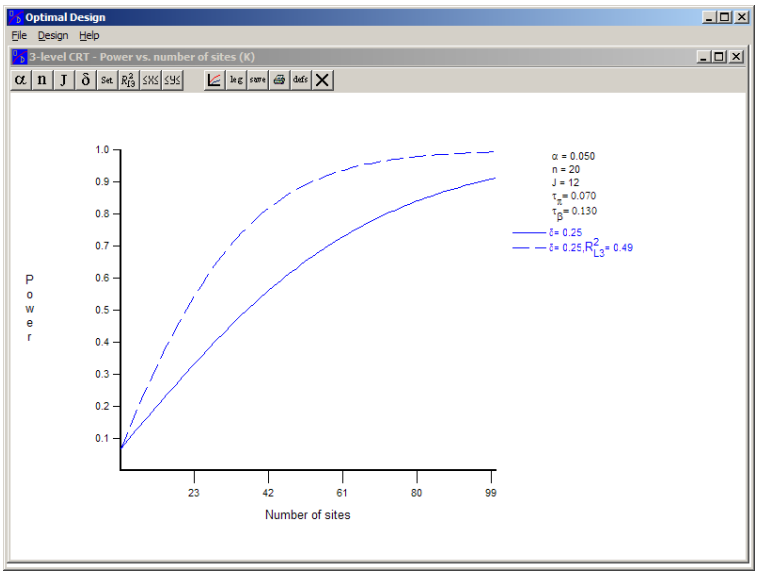


Figure 8.4. Two power curves.

According to the key, the dotted trajectory is the power curve when the covariate is included. In this case, approximately 40 total schools are needed, a reduction of 32 schools, which may greatly reduce the cost of the study.

The example provided in this section placed the sample size on the x-axis. However, any of the other parameters could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

8.9 Effect size approach for conducting a power analysis

The researchers conducting the study are limited to 60 schools, 30 in the treatment and 30 in the control group with 20 students per class. Based on past studies, they expect about 13 percent of the variation to lie between schools and about 7 percent of the variation to lie between classrooms within schools. What is the MDES the researchers can find with power = 0.80? Assuming the school-level covariate explains 49% of the variation in test scores, what is the MDES?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. total number of sites (*K*). This allows the user to see how the MDES changes as a function of the power holding all other parameters constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 3 → Power on y-axis (continuous outcome) → MDES vs. number of sites (*K*) as shown in Figure 8.5.

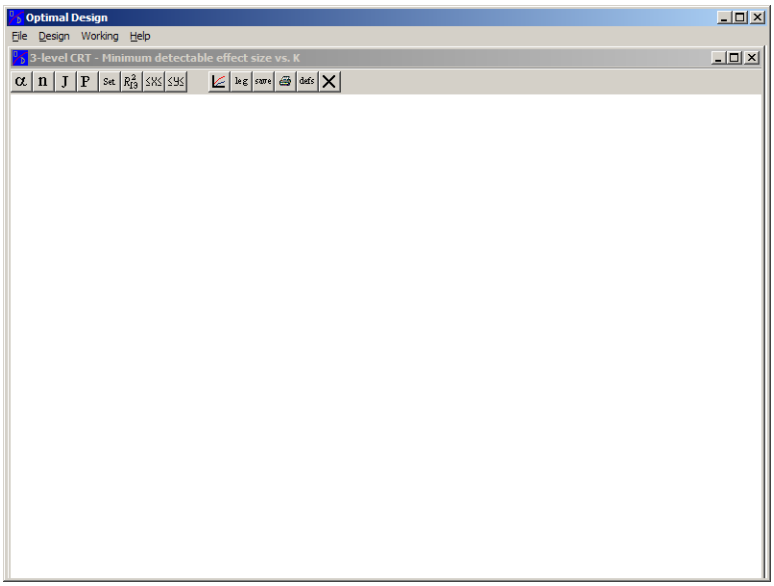


Figure 8.5. Initial blank screen of MDES vs. number of sites (*K*)

Step 2: Click on n. Set n(1) = 25. By clicking on n(1) = 25, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J* (1) = 12.

Step 4: Click on *P*. Set *P* (1) = 0.80.

Step 5: Click on set. Set ρ_{π} = 0.07 and ρ_{β} = 0.13. The resulting power curve appears in Figure 8.6.

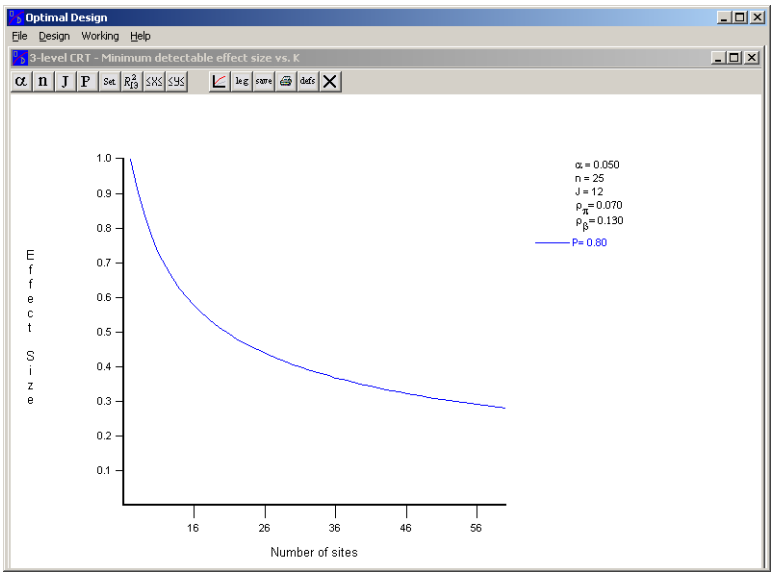


Figure 8.6. Power curve.

Clicking along the trajectory we can see that for K = 30, the MDES is approximately 0.40. Let's see what happens when we add a cluster-level covariate.

Step 6: Click on R. Set R^2 – 2 = 0.49. Figure 8.7 displays the two power curves.

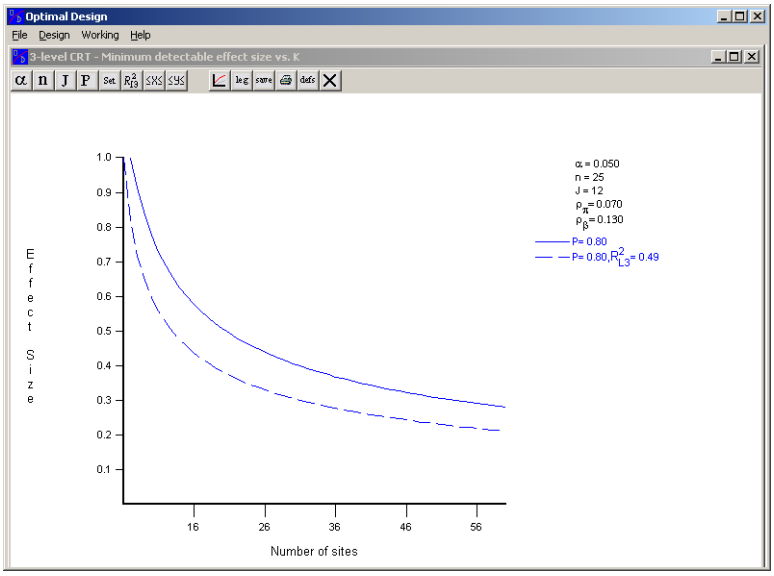


Figure 8.7. MDES vs. power with a covariate.

Clicking along the dotted trajectory, the MDES with K = 30 is 0.31, reducing the MDES by 0.09 effect size units.

The examples in this section are meant to provide at guide to users for how to use the 3-level CRT. We described Power vs. number of sites (K) and MDES vs. number of sites (K). The other options function similarly, and simply place a different parameter on the x-axis. The choice

of which module is most appropriate depends on the unknown parameters. However, all modules yield the same results if identical parameters are used so the choice depends on what module is most closely aligned with the known and unknown parameters in a study.

9.0 Multi-site Cluster Randomized Trials with Treatment at Level 2

A design using blocking before randomizing groups can be thought of as a multisite cluster randomized trial (MSCRT), an extension of the cluster randomized trial. In a MSCRT, the site is the block and clusters are randomly assigned to treatment and control within each site. Sometimes the sites are natural administrative units, for example, schools where classrooms are randomly assigned to treatment within schools. Sometimes sites are formed by the researcher by creating blocks of units that are similar. For example, schools may be matched according to percent of students who receive free/reduced lunch. Within each match, one school is randomly assigned to the treatment condition and one school to the control condition. Students are nested within schools. As discussed in Section I, pre-randomization blocking is often employed to increase the precision of the estimate and the power of the study and/or to improve the face validity of the study.

The choice of whether to treat the site effects as random or fixed affects the power. We discuss the models assuming the sites are random effects first followed by the models assuming fixed site effects.

9.1 The model (assuming random site effects)

We can represent data from a multi-site cluster randomized trial as a three level model, persons nested within clusters nested within sites. The level-1 model, or person-level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \qquad [9.1]$$

for $i \in \{1,2,...,n\}$ persons per cluster, $j \in \{1,2,...,J\}$ clusters and $k \in \{1,2,...,K\}$ sites,

where π_{0jk} is the mean for cluster *j* in site *k*;

e_{ijk} is the error associated with each person; and

σ^2 is the within-cluster variance.

The level-2 model, or cluster-level model, is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} W_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi) \qquad [9.2]$$

where β_{00k} is the mean for site *k*;

β_{01k} is the treatment effect at site *k*;

W_{jk} is a treatment contrast indicator, ½ for treatment and -½ for the control;

r_{0jk} is the random effect associated with each cluster; and

τ_{π} is the variance between clusters within sites.

The level-3 model, or site-level model, is:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad \text{var}\ (u_{00k}) \sim \tau_{\beta_{00}}$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad \text{var}\ (u_{01k}) \sim \tau_{\beta_{11}} \qquad \text{cov}(u_{00k}, u_{01k}) = \tau_{\beta_{01}} \quad [9.3]$$

where γ_{000} is the grand mean;

γ_{010} is the average treatment effect ("main effect of treatment");

u_{00k} is the random effect associated with each site mean;

u_{01k} is the random effect associated with each site treatment effect;

$\tau_{\beta_{00}}$ is the variance between site means;

$\tau_{\beta_{11}}$ is the variance between sites on the treatment effect; and

$\tau_{\beta_{01}}$ is the covariance between site-specific means and site-specific treatment effects.

The random effects u_{00k} and u_{01k} are typically assumed bivariate normal in distribution. We are interested in two quantities, the main effect of treatment, γ_{010}, and the variance of the treatment effect, $\tau_{\beta_{11}}$. Note that we are operating under a random effects model. In a fixed effects model, the variance of the treatment effect, $\tau_{\beta_{11}}$, would be 0.

9.2 Testing the treatment effect

The average treatment effect is denoted as γ_{010} in level 3 of the model. Given a balanced design, it is estimated by

$$\hat{\gamma}_{010} = \bar{Y}_E - \bar{Y}_C \qquad [9.4]$$

where $\bar{Y}_E$ is the mean for the experimental group and $\bar{Y}_C$ is the mean for the control group.

Note that the estimated main effect of treatment looks like that in the cluster randomized trial except that now we are summing over clusters and sites. Thus the variance of the treatment effect is slightly different than in a cluster randomized trial. It is estimated by (Raudenbush & Liu, 2000)

$$Var(\hat{\gamma}_{010}) = \frac{\tau_{\beta_{11}} + 4(\tau_\pi + \sigma^2 / n)/J}{K}. \qquad [9.5]$$

The main difference between the variance of the treatment effect in a multi-site cluster randomized trial and that in a cluster randomized trial is that we now have four sources of variability, the within-cluster variance, σ^2, the between-cluster variance or within-site variance, τ_π, the between-site variance, $\tau_{\beta_{00}}$, and the between-site variance in the treatment effect, $\tau_{\beta_{11}}$.

If the data are balanced, we can use the results of a nested analysis of variance with random effects for the clusters and sites and fixed effects for the treatment. Similar to prior tests, the test statistic is an *F* statistic. The *F* test follows a non-central *F* distribution, $F(1, K\text{-}1; \lambda)$. Recall that the noncentrality parameter is a ratio of the squared-treatment effect to the variance of the treatment effect estimate. Below is the noncentrality parameter for the test.

$$\lambda = \frac{\gamma_{010}^2}{\text{var}(\hat{\gamma}_{010})} = \frac{K\gamma_{010}^2}{\tau_{\beta_{11}} + 4(\tau_\pi + \sigma^2 / n)/J}. \qquad [9.6]$$

Recall that the larger the non-centrality parameter, the greater the power of the test. By looking at the formula, we can see that *K*, the number of sites, has the greatest impact on the power. It is especially important to have a large *K* if there is a lot of between-site variance. Increasing *J* also increases the power but is not as important as *K*. *J* becomes more important if there is a lot of variability between clusters. Finally, increasing *n* does increase the power, but has the smallest effect of the three sample sizes. Increasing *n* is most beneficial if there is a lot of variability within clusters. In addition to *K*, *J*, and *n*, a larger effect size increases power. Note that $\tau_{\beta_{11}}$, the between-site variance of the treatment effect, appears in the denominator of the non-centrality parameter. As mentioned above, if the variance of the treatment effect across sites is large, it is particularly important to have a large number of sites to counteract the increase in variance in order to achieve adequate power. However, if the variability of the impact across sites is very large, the average treatment effect may not be informative.

Thus far, we have focused on the unstandardized random effects model for a multi-site cluster randomized trial. In a multi-site cluster randomized trial, we need to think about the standardized effect size and the effect size variability. The magnitude of the effect size variability depends on the desired effect size. For example, an effect size variance of 0.10 is the same as a standard error of approximately $\sqrt{0.10} = 0.31$. If a researcher desires a minimum

detectable effect of 0.20, a standard error of 0.31 seems too large and would indicate a lot of uncertainty in the estimate. For an effect size of 0.20, an effect size variance of 0.01 (or standard error of 0.10) seems more reasonable. Section 5.3 presents the standardized notation.

9.3 Standardized notation

In standardized notation, the non-centrality parameter, λ, can be rewritten as:

$$\lambda = \frac{K\gamma_{010}^2/(\tau_\pi + \sigma^2)}{\tau_{\beta_{11}} + 4(\tau_\pi + \sigma^2/n)/J(\tau_\pi + \sigma^2)} = \frac{K\delta^2}{\sigma_\delta^2 + 4[\rho + (1-\rho)/n]/J} \qquad [9.7]$$

where the intra-cluster correlation, ρ, is:

$$\rho = \frac{\tau_\pi}{\tau_\pi + \sigma^2},$$

or the variance between clusters relative to the between and within cluster variation within blocks; δ is the standardized main effect of treatment,

$$\delta = \frac{\gamma_{010}}{\sqrt{\tau_\pi + \sigma^2}}$$

and σ_δ^2 is the variance of the standardized treatment effect**,**

$$\sigma_\delta^2 = \frac{\tau_{\beta_{11}}}{\tau_\pi + \sigma^2}.$$[5]

9.4 Using a covariate to increase power

In addition to blocking, researchers may also have cluster-level covariates available. The cluster-level covariate in a multi-site randomized trial functions similarly to the cluster-level covariate in a cluster randomized trial. Recall that including a cluster-level covariate influences the power of the test depending on the strength of the correlation between the covariate and the true cluster mean outcome, or how much of the variability in the true cluster mean outcome is explained by the covariate. The proportion of explained variability is denoted $\rho_{x\pi_0}^2$. The larger $\rho_{x\pi_0}^2$, the smaller the conditional level 2 variance, $\tau_{\pi|x}$, relative to the unconditional level 2 variance, τ_π, and the greater the benefit of the covariate in increasing precision and power.

9.5 The model with a cluster-level covariate

[5] The Optimal Design asks for slightly different parameters than those presented in the non-centrality parameter in equation 7. The Optimal Design asks to user to enter the between cluster variance prior to blocking and the percent of variance explained by blocking. It calculates the parameters in equation 9.7 within the program.

The level 1 model for a multi-site cluster randomized trial with a cluster-level covariate looks the same as the level-1 model for a regular multi-site cluster randomized trial (see equation 9.1). The level 2 model looks slightly different because it includes the cluster level covariate. It is written as:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} W_{jk} + \beta_{02k} X_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_{\pi|x}) \qquad [9.8]$$

$$\text{Note: } \tau_{\pi|x} = (1 - \rho^2_{x\pi_0})\tau_\pi$$

where β_{00k} is the adjusted mean for site *k*;

β_{01k} is the adjusted treatment effect at site *k*;

β_{02k} is the regression coefficient for the cluster-level covariate at site *k*;

W_{jk} is 0.5 for treatment and –0.5 for control;

X_{jk} is the cluster level covariate, typically centered to have mean 0;

r_{0jk} is the random effect associated with each cluster; and

$\tau_{\pi|x}$ is the residual variance conditional on the cluster-level covariate X_{jk}.

Note that the between cluster variance is now a residual variance conditional on the cluster-level covariate X_{jk}.

The level 3 model is now:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}|x}) \qquad [9.9]$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad u_{01k} \sim N(0, \tau_{\beta_{11}})$$

$$\beta_{02k} = \gamma_{020}$$

where γ_{000} is the grand mean;

γ_{010} is the average treatment effect ("main effect of treatment");

γ_{020} is the regression coefficient for the cluster-level covariate, which is assumed constant across sites;

u_{00k} is the random effect associated with each site mean;

u_{01k} is the random effect associated with each site treatment effect;

$\tau_{\beta_{00}|x}$ is the residual variance between site means; and

$\tau_{\beta_{11}}$ is the variance between sites on the treatment effect.

Because of the randomization, the true treatment effect is not influenced by the covariate. Thus it is not necessary to have a conditional variance for the between-site variation in the treatment effect. Note that we are also fixing the average regression coefficient for the cluster-level covariate.

9.6 Testing the treatment effect (Including a cluster-level covariate)

The estimate of the main effect of the treatment accounting for the cluster-level covariate is:

$$\hat{\gamma}_{010} = \bar{Y}_E - \bar{Y}_C - \hat{\gamma}_{020}(\bar{X}_E - \bar{X}_C)\,. \qquad [9.10]$$

In words, it is the mean difference adjusted for the treatment group differences on the covariate. To test the main effect of treatment we use an F-statistic which follows a non-central *F* distribution, $F(1, K-1; \lambda_x)$ where:

$$\lambda_x = \frac{K\gamma_{010}^2}{\tau_{\beta_{11}} + 4(\tau_{\pi|x} + \sigma^2/n)/J} \qquad [9.11]$$

This formula for the noncentrality parameter looks similar to the noncentrality parameter without the covariate except that the estimate of the treatment effect is calculated differently and the between cluster variance is now a conditional variance.

Following the same logic as the multi-site cluster randomized trial with no covariate, we can standardize the parameters. The non-centrality parameter expressed in standardized notation is:

$$\lambda_x = \frac{K\delta^{*2}}{\sigma_\delta^{2*} + 4\left[\rho^* + (1-\rho^*)/n\right]/J} \qquad [9.12]$$

where ρ^* the intra-cluster correlation,

$$\rho^* = \frac{\tau_{\pi|x}}{\tau_{\pi|x} + \sigma^2}\,,$$

or the conditional variance between clusters relative to the between and within cluster variation within blocks; δ^* is the standardized main effect of treatment conditional on the covariate,

$$\delta^* = \frac{\gamma_{010}}{\sqrt{\tau_{\pi|x} + \sigma^2}}\,;$$

and σ_{δ}^{2*} is the variance of the standardized treatment effect conditional on the covariate,

$$\sigma_{\delta}^{2*} = \frac{\tau_{\beta_{11}}}{\tau_{\pi|x} + \sigma^2}.$$

Because the conditional standardized quantities resulting from inclusion of a covariate are frequently unknown, the program asks the user to enter the unconditional parameters, ρ, δ, and σ_{δ}^2. The program calculates the conditional standardized values based on the input.

9.7 Testing the variance of the treatment effect

Recall that in a fixed effects model we assume the treatment effect to be homogeneous across the sites. Thus the tests described in this section are only applicable under the random effects model where we assume the treatment effect randomly varies across the sites. To quantify this difference, we estimate the variance of the treatment effect across the sites. The design, with treatments randomized to clusters within sites, allows us to estimate this variability. If it is very large, it may be hiding the true treatment effect. For example, imagine a multi-site cluster randomized trial that reports a treatment effect of 0.23. The researchers claim that the new reading program improves scores by 0.23 units. However, they fail to report that the standardized treatment effect variability across sites is 0.30. The high variance suggests that some types of schools benefit from the program while other types of schools actually suffer from the program. For example, there may be a differential effect by location, where rural schools that adopt the program see positive effects but urban schools that adopt the program see negative effects. Thus the researchers would need to investigate moderating site characteristics. Reporting the average treatment effect alone may be very misleading and is not recommended.

Because the variance of the treatment effect is critical in determining the interpretation of a treatment effect estimate, it is important to be able to detect the treatment effect variability with adequate power. The remainder of this section describes how to calculate the power for the variance of the treatment effect using standardized notation.

The null and alternative tests for the treatment effect variability are:

$H_0 : \sigma_{\delta}^2 = 0$

$H_1 : \sigma_{\delta}^2 > 0$.

The null hypothesis states that the variance of the treatment impact across sites is 0, whereas the alternative hypothesis states that it is greater than 0. The test for the variance of the treatment effect is an *F* test. The *F* statistic (Raudenbush & Liu, 2000):

$$F = \frac{\hat{\tau}_{\beta_{11}} + 4(\hat{\tau}_{\pi} + \frac{\hat{\sigma}^2}{n}) / J}{4(\hat{\tau}_{\pi} + \frac{\hat{\sigma}^2}{n}) / J}. \qquad [9.13]$$

Note that the average effect size is not a part of the calculation, thus the power is based on the number of sites, *K*, the number of clusters per site, *J*, the number of people per cluster, *n*, the effect size variability, $\tau_{\beta_{11}}$ and the intra-cluster correlation, ρ. The *F* statistic follows a central *F* distribution with df = *K*-1, *K*(*J*-2). The ratio of the expectation of the numerator to the expectation of the denominator, in standardized notation, is

$$\omega = 1 + \frac{J\sigma_{\delta}^2}{4[\rho + (1-\rho)/n]}. \qquad [9.14]$$

Under the null hypothesis, we expect σ_{δ}^2 to be 0, thus $\omega = 1$. As σ_{δ}^2 increases, ω gets larger, increasing the power of the test. Thus the number of clusters within each site is critical for increasing the power to detect the variance of the treatment effect across sites. As the number of clusters within each site increases, so does the power to detect the variability of treatment effects. Increasing *K* also increases the power, through the degrees of freedom, but is not as important as increasing *J*. Note that this is the opposite of what we found in the case of power for the treatment effect, where *K* is the most significant factor in increasing power and *J* is less important.

Looking at equation 9.14, we can see that it will be difficult to achieve adequate power to detect small values of σ_{δ}^2, like 0.01 unless *J* is extremely large, which is unlikely. This is not a major problem because our primary concern is to be able to detect larger treatment effect variability since small values will not influence the interpretation of the treatment effect.

9.8 The fixed effects model

The fixed effects model is identical to the random effects model with a crucial exception: the site-specific contributions u_{00k} and u_{01k} are designated as fixed constants rather than random variables.

The level-1 and level-2 models are identical to equations 9.1 and 9.2 in the random effects case. The level-3 model, or site-level model is:

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad [9.15]$$

where γ_{000} is the grand mean;

γ_{010} is the average treatment effect ("main effect of treatment");

u_{00k}, for $k \in \{1,2,...,K\}$, are fixed effects associated with each site mean, constrained to have a mean of zero; and

u_{01k}, for $k \in \{1,2,...,K\}$, are fixed effects associated with each treatment-by-site interaction, constrained to have a mean of zero.

We are interested in two kinds of quantities, the main effect of treatment, γ_{010}, and the fixed treatment-by-site interaction effects u_{01k}, for $k \in \{1,2,...,K\}$.

9.9 Testing the average treatment effect

We can use the results of a nested analysis of variance with random effects for the clusters and fixed effects for sites, treatments, and site-by-treatment interaction. Similar to prior tests, the test statistic is an *F* statistic. The *F* test follows a non-central *F* distribution, $F(1, K(J-2); \lambda)$. Recall that the noncentrality parameter is a ratio of the squared-treatment effect to the variance of the treatment effect estimate. Below is the noncentrality parameter for the test.

$$\lambda = \frac{KJ\gamma_{010}^2}{4(\tau_\pi + \sigma^2 / n)} \ . \qquad [9.16]$$

Recall that the larger the non-centrality parameter, the greater the power of the test. By looking at the formula, we can see that *KJ*, the total number of clusters, has the greatest impact on the power. Finally, increasing *n* does increase the power, but has the smallest effect of the three sample sizes. Increasing *n* is most beneficial if there is a lot of variability within clusters. In addition to *K*, *J*, and *n*, a larger effect size increases power. Note that unlike the case of the random effects model, $\tau_{\beta_{11}}$, the variance of the treatment effect, does not appear in the denominator of the non-centrality parameter. However, if the variation of the treatment effects across sites is large, the average treatment effect may not be informative because it may not characterize the treatment effect in any given site. Section 9.10 discusses the test of the variation

site by treatment effect variation under the fixed effects model. If the treatment effects vary across sites with a fixed effects model, the main effect of treatment is interpreted with great caution.

In the fixed effects standardized model, the non-centrality parameter, λ, can be rewritten in terms of the standardized model:

$$\lambda = \frac{KJ\delta^2}{4[\rho + (1-\rho)/n]} \qquad \text{(see footnote 5)} \qquad [9.17]$$

where ρ is the intra-cluster correlation ,

$$\rho = \frac{\tau_\pi}{\tau_\pi + \sigma^2},$$

or the variance between clusters relative to the between and within cluster variation within blocks; and δ is the standardized main effect of treatment,

$$\delta = \frac{\gamma_{010}}{\sqrt{\tau_\pi + \sigma^2}}.$$

9.10 Testing site-by-treatment variation in the context of a fixed effects model.

Operationally, the test of the site-by-treatment variation in the case of the fixed effects model is identical to that in the case of the random effects model (see Section 6.6 "Testing the Variance of the Treatment Effect"). The null hypothesis, however, differs. Recall that in the case of the random effects model we test

$$H_0 : \tau_{\beta 11} = 0$$

or for the standardized random effects model, we test

$$A_0 : \sigma_\delta^2 = 0.$$

However, in the fixed effect model, the site-specific treatment effects are fixed constants rather than random variables. Thus we have, in the non-standardized model

$$H_0 : \sum_{k-1}^{K} u_{01k}^2 = 0.$$

As in the random effects case, we test this hypothesis using

$$F[K-1, K(J-2)] = \frac{MS\ treatments\ \ by\ site}{MS\ within\ \ cell}.$$

When the *F* test indicates rejection of H_0, one emphasizes the estimation of site-specific treatment effects (also known as "simple main effects" – see Kirk (1982), p. 365) or post hoc procedures designed to identify subsets of sites for which the treatment effect is homogeneous (see Kirk (1982), p. 317).

9.11 Using Optimal Design for the multi-site cluster randomized trials (MSCRT)

The menu for the MSCRT is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 2. In this chapter we focus on continuous outcomes thus the first two options are shown below.

Power on y-axis (continuous outcomes)

- Power vs. cluster size (*n*)
- Power vs. number of sites (*K*)
- Power vs. number of clusters per site (*J*)
- Power vs. intraclass correlation (ρ)
- Power vs. effect size (delta)
- Power vs. effect size variability
- Power vs. proportion of explained variation by level-two covariate (R2)

MDES on y-axis (continuous outcomes)

- MDES vs. cluster size (*n*)
- MDES vs. number of sites (*K*)
- MDES vs. number of cluster per site (*J*)
- MDES vs. intraclass correlation (ρ)
- MDES vs. power (*P*)
- MDES vs. effect size variability
- MDES vs. proportion of explained variation by level-two covariate (R2)

The first set of options present the power on the y-axis and either the cluster size, number of clusters per site, number of sites, interclass correlation, effect size, effect size variability, or proportion of explained variation by level-2 covariate to vary on the x-axis. The second set of options present the MDES on the y-axis and either the cluster size, number of clusters per site, number of sites, intraclass correlation, power, effect size variability, or proportion of explained

variation by level-2 covariate to vary on the x-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

9.12 Example

Suppose a team of researchers develop a new literacy program. The founders of the new program propose that students who participate in the program will have increased reading achievement. They propose a three-level design with students nested within schools within districts and they want to block by district. That is, within each district, half of the schools will be randomly assigned to the new program and half to the current program. They plan to test students who are in classrooms that participate in the new program (experimental group) and students who are in classrooms that participate in the regular program (control group) in each of the schools using a reading test to determine if students using the new program score higher.

9.13 Power determination approach for conducting a power analysis

Based on past studies, the researchers expect about 25 percent of the variation in the outcome to be between schools (prior to blocking). By blocking on district, the researchers expect to explain 40% of the variation in the outcome variable. They are interested in detecting an effect size of at least 0.25 with adequate power. Assuming that they plan to test 200 students per school and have secured 10 schools per district, how many districts (sites) are necessary to achieve power = 0.80? Assume the researchers are interested in generalizing to a broader population of districts thus they treat the districts as random effects and assume an effect size variability of 0.01. Suppose the researchers have access to a school level pre-test that explain about 49% of the variation in post-test scores. How many districts are required after including the school-level covariate?

In Scenario 1, the number of sites, *K*, is unknown. As a result, we want to select the power vs. number of sites (*K*) option. This allows the number of sites to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 2 → Power on for treatment effect on y-axis (continuous outcome) → Power vs. number of sites (*K*) as shown in Figure 9.1.

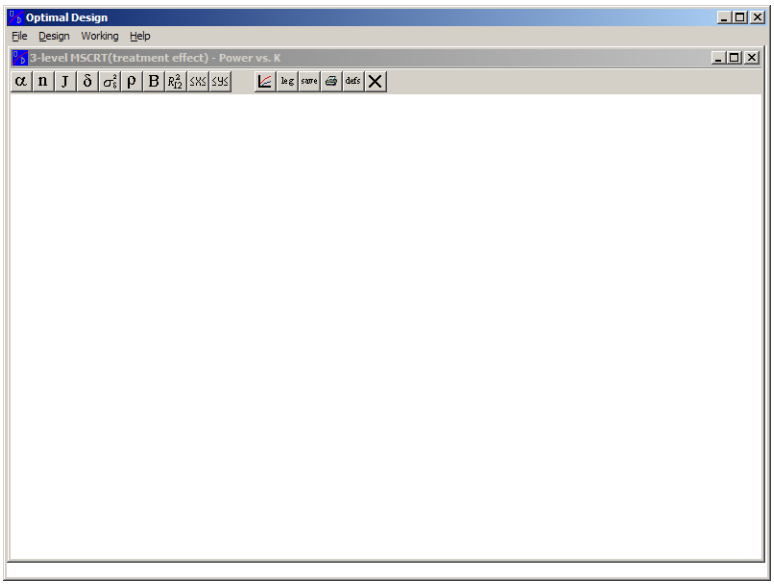


Figure 9.1. Initial screen for cluster randomized trials with blocking.

The toolbar at the top includes the parameters required for calculating the power: sample size within cluster (n), number of clusters per site (*J*), effect size (δ), effect size variability (σ_δ^2), intraclass correlation, (ρ), percent of variance explained by blocking, (B), and explained proportion of variance by covariate (R). The number of sites (*K*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 200. By clicking on n(1) = 200, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J*(1) = 10. This is the number of schools per district.

Step 4: Click on δ. Set delta (1) = 0.25.

Step 5: Click on σ_δ^2. Set sigma2d = 0.01. Setting the effect size variability greater than 0 assumes random site effects.

Step 6: Click on ρ. Set rho (1) = 0.25.

Step 7: Click on B. Set B(1) = 0.40. This defines the percent of variance explained by blocking on district. The resulting power curve appears in Figure 9.2.

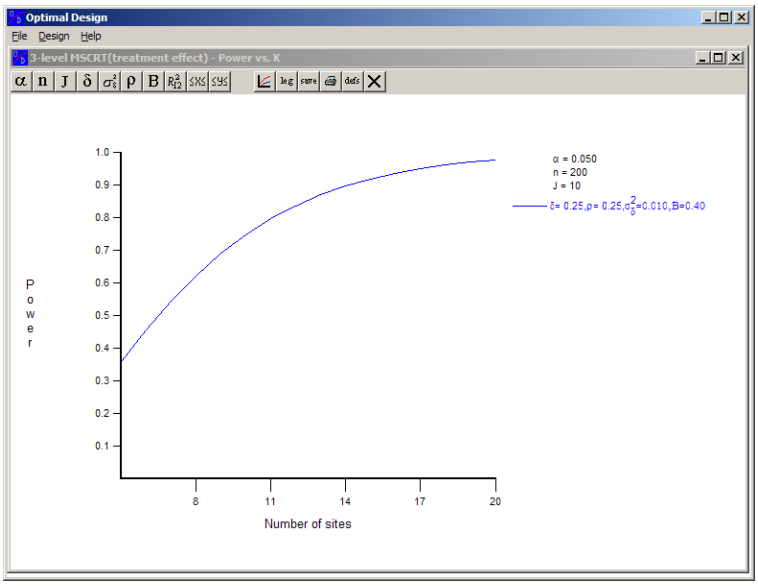


Figure 9.2. Power curve.

Clicking along the power curve, we can see that 12 districts are required for the study. Within each district, we randomly assign 5 schools to the treatment condition and 5 schools to the control condition. Let's see what happens when we include the school-level covariate.

Step 8: Click on R. Set R2 – 2 equal to 0.49. Two power curves appear as shown in Figure 9.3.

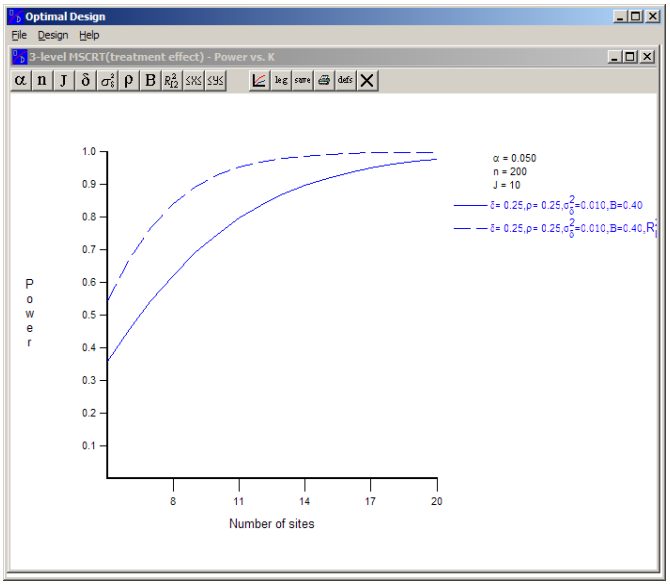


Figure 9.3. Power curves.

According to the key, the dotted trajectory is the power curve when the covariate is included. In this case, approximately 8 districts are needed, a reduction of 4 districts and 40 schools, which may greatly reduce the cost of the study.

If we were not interested in generalizing to a larger population of districts, we could consider the districts as fixed effects. In this case, we would set $\sigma_\delta^2 = 0$.

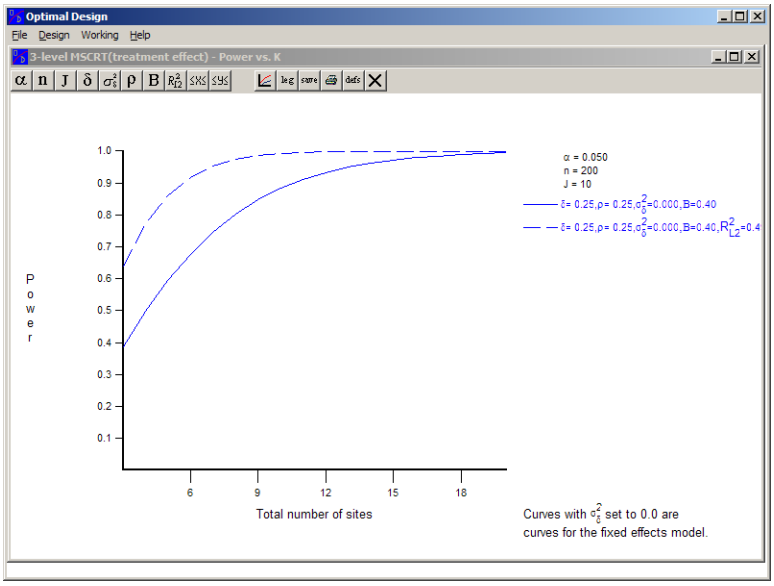


Figure 9.4. Fixed effects power curve.

Clicking on the curve reveals that 5 or 8 sites, with and without a covariate, are required for power = 0.80. The number of required sites decreases when the sites are treated as fixed effects.

The example provided in this section placed the number of districts on the x-axis. However, any of the other parameters could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

9.14 Effect size approach for conducting a power analysis

Based on past studies, the researchers expect about 25 percent of the variation in the outcome to be between schools (prior to blocking). By blocking on district, the researchers expect to explain 40% of the variation in the outcome variable. The researchers have secured 8 districts (sites), 10 schools per district, and 200 students per school. What is the MDES? Assume the researchers are interested in generalizing to a broader population of districts thus they treat the districts as random effects and assume an effect size variability of 0.01. Suppose the researchers have access to a school level pre-test that explain about 49% of the variation in post-test scores. What is the MDES after including the covariate?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. total number of sites (*K*). This allows the user to see how the MDES changes as a function of the power holding all other parameters

constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 2 → MDES for treatment effect on y-axis (continuous outcome) → MDES vs. total number of sites (*K*) as shown in Figure 9.5.

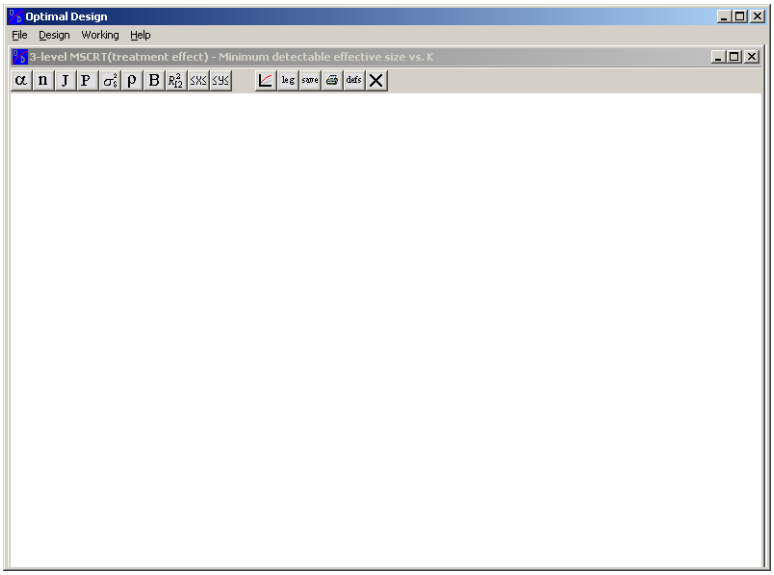


Figure 9.5. Initial blank screen of MDES vs. total number of sites (*K*)

Step 2: Click on n. Set n(1) = 200. By clicking on n(1) = 200, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J* (1) = 10.

Step 4: Click on *P*. Set *P* (1) = 0.80.

Step 5: Click on σ^2_δ. Set sigma = 0.01

Step 6:. Click on ρ. Set ρ=0.25.

Step 7: Click on B. Set B = 0.40. The resulting power curve appears in Figure 9.6.

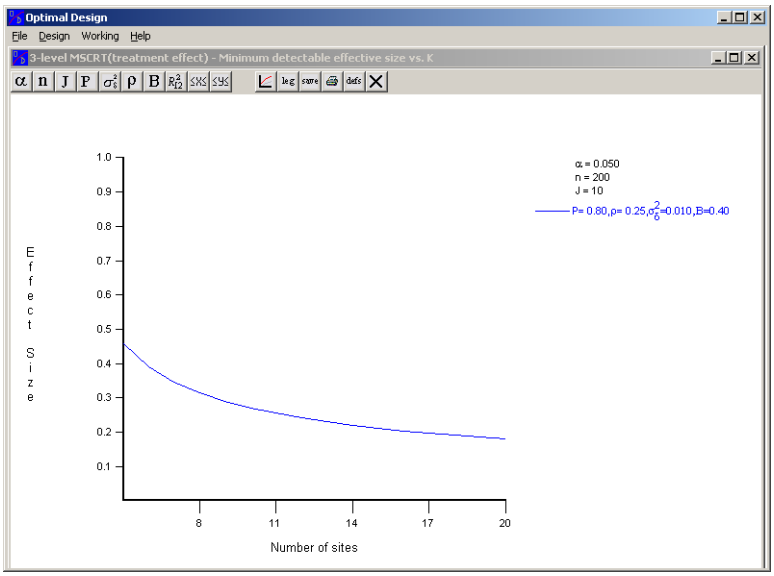


Figure 9.6. Power curve with specified parameters.

Clicking along the trajectory we can see that for K = 8, the MDES is approximately 0.32. If the literature suggests that an effect size of 0.25 is more likely, clicking along the curve for an effect size of 0.25 reveals that the study will need 12 sites. Let's see what happens when we add a cluster-level covariate.

Step 8: Click on R. Set R – 2 = 0.49. Figure 9.7 displays the two power curves.

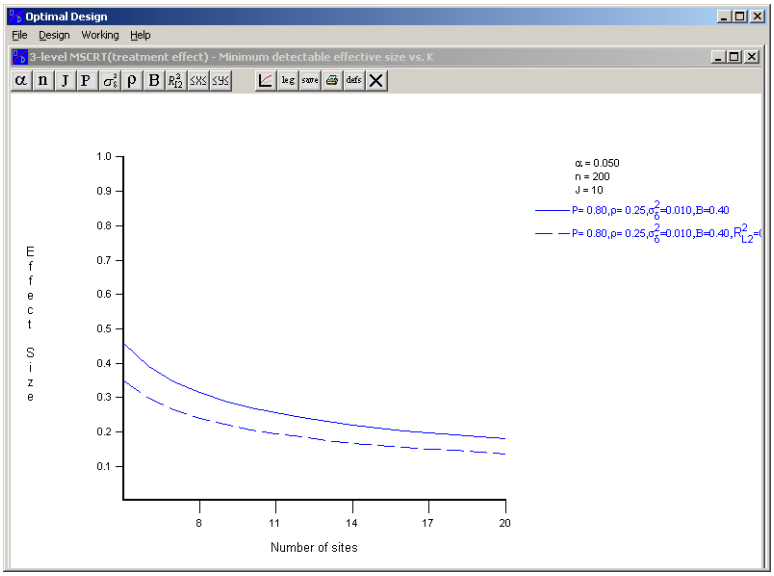


Figure 9.7. Power curve with covariate.

Clicking along the dotted trajectory, the MDES with *K* = 8 is 0.24, reducing the MDES by 0.07 effect size units.

9.15 Power for effect size variability

Thus far we have focused on power calculations for the treatment effect in a random effects model. Researchers may also be interested in the power for effect size variability. Recall that if the effect size variability is large, the treatment effect may be meaningless and it is important to investigate moderating effects to explain the variability in effect sizes. As a result, it is important to be able to detect the effect size variability with adequate power. For example, recall the literacy program. The researchers propose a blocked design with students nested within schools which are blocked by district. They expect 8 districts, 10 schools per district, and 200 students per school. They expect blocking to explain 40% of the variability in the outcome. They plan to test students who are in schools that participate in the regular program (control group) and students who are in classrooms that participate in the new program (experimental group) in each of the participating schools using a reading test. They want to know the power to detect the variability in the effect sizes across sites. Assuming an intraclass correlation of 0.25 and explained variation by covariate of 0.49, what power do the researchers have to detect an effect size variability of 0.01? of 0.05?

Step 1: Select Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 2 → Power for effect size variability on y-axis (continuous outcome) → Power vs. effect size variability. The blank screen is in Figure 9.8.

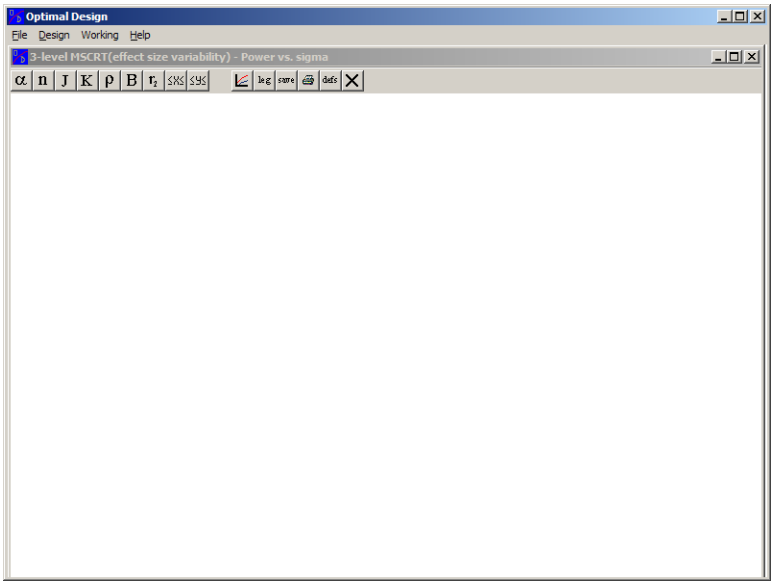


Figure 9.8. Initial blank screen for power vs. effect size.

The primary difference on the toolbar is that there is no effect size, δ, because the effect size is not a part of the calculation.

Step 2: Click on n. Set n(1) = 200.

Step 3: Click on *J*. Set *J*(1) = 10.

Step 4: Click on *K*. Set *K*(1) = 8.

Step 5: Click on ρ. Set rho (1) = 0.25.

Step 6: Click on B. Set B = 0.40.

Step 7: Click on R. Set R = 0.49. The final curves are in Figure 9.9.

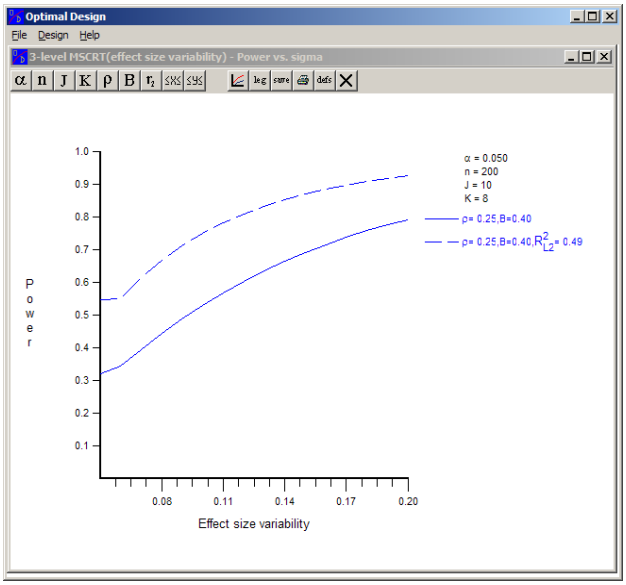


Figure 9.9. Power to detect an effect size with variability.

Figure 9.9 reveals that the power to detect an effect size variability of 0.01 is less than 0.05. We can change the scale of the x-axis to get the exact power.

Step 8: Click on <X<. Set minimum = 0.001. The new curves are in Figure 9.10.

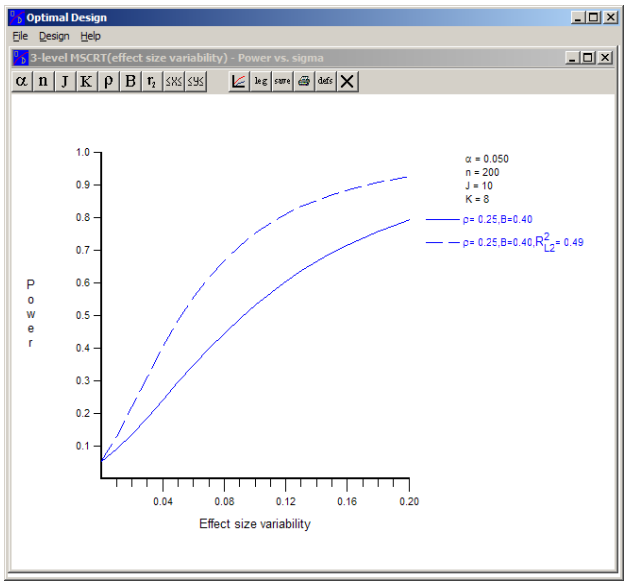


Figure 9.10. Power curve.

Clicking along the dotted trajectory reveals power of approximately 0.12 for an effect size variability of 0.01. Although the power is very small, it is not particularly problematic because an effect size variability of 0.01 is so small it is unlikely to change the meaning of the treatment effect. The power to detect an effect size variability of 0.10 is about 0.77, which is much higher and important because an effect size variability of 0.10 could mask the true treatment effect.

10.0 Multi-site cluster randomized trials with Treatment at Level 3

A trial with 4 levels in which level 4 represents the blocks or sites can be thought of as a multi-site cluster randomized trials with 4 levels. For example, suppose that students are nested within classrooms, nested within schools, nested within districts. Within each district, the schools are randomly assigned to either the treatment or control. In other words, the districts are blocks so this is a multisite cluster randomized trial with a total of four levels. This chapter focuses on power for the treatment effect. We distinguish between fixed and random site effects and begin with the models for treating sites as random effects. Although the variance of the treatment effect is important, the tests for this are not included in this chapter or in the current version of Optimal Design.

10.1The model (assuming random site effects)

Data from a four-level multi-site cluster randomized trial can be written as a four level hierarchical model. The level-1 model is:

$$Y_{ijkl} = \pi_{0\,jkl} + e_{ijkl} \qquad e_{ijkl} \sim N(0, \sigma^2) \qquad [10.1]$$

for $i \in \{1,2,...,n\}$ persons per classroom, $j \in \{1,2,...,J\}$ classrooms per school, $k \in \{1,2,...,K\}$ schools per site, and $l \in \{1,2,...,L\}$ sites,

where $\pi_{0\,jkl}$ is the mean for classroom *j* in school *k* in site *l;*

e_{ijkl} is the error associated with each person; and

σ^2 is the within-classroom variance.

The level-2 model is:

$$\pi_{0\,jkl} = \beta_{00kl} + r_{0\,jkl} \qquad r_{0\,jkl} \sim N(0, \tau_\pi) \qquad [10.2]$$

where β_{00kl} is the mean for school *k* in site *l;*

$r_{0\,jkl}$ is the random effect associated with each classroom; and

τ_π is the variance between classrooms within schools.

The level-3 model with a school-level covariate is:

$$\beta_{00kl} = \gamma_{000l} + \gamma_{001l} W_{00kl} + u_{00kl} \qquad u_{00kl} \sim N(0, \tau_\beta) \qquad [10.3]$$

where γ_{000l} is the mean for school l;

γ_{001l} is the treatment effect at school l;

W_{00kl} is the indicator variable where -1/2 represents the control group and ½ represents the treatment group;

u_{00kl} is the random effect associated with school; and

τ_{β} is the variance between schools within sites.

The level-4 model is:

$$\begin{aligned} \gamma_{000l} &= \eta_{0000} + s_{000l} \\ \gamma_{001l} &= \eta_{0010} + s_{001l} \end{aligned} \qquad \begin{aligned} &\text{var}(s_{000l}) \sim N(0, \tau_{\gamma_{000}}) \\ &\text{var}(s_{001l}) \sim N(0, \tau_{\gamma_{011}}) \end{aligned} \qquad [10.4]$$

where η_{0000} is the grand mean;

η_{0010} is the average treatment effect;

s_{000} is the random effect associated with each site mean;

s_{001} is the random effect associated with each site treatment effect;

$\tau_{\gamma 000}$ is the variance between site means;

$\tau_{\gamma 111}$ is the variance between sites on the treatment effect; and

$\tau_{\gamma 001}$ is the covariance between site-specific means and site-specific treatment effects.

The two quantities of interest are the main effect of treatment, η_{0010}, and the variance between the sites on the treatment effect, $\tau_{\gamma 111}$. The variance of the treatment effect functions similarly to the case of the three-level multi-site cluster randomized trial and is not discussed here.

10.2Testing the treatment effect

Given a balanced design, the average treatment effect is estimated by:

$$\hat{\eta}_{0010} = \bar{Y}_E - \bar{Y}_C \qquad [10.5]$$

where $\bar{Y}_E$ is the mean for the experimental group and $\bar{Y}_C$ is the mean for the control group. The variance of the treatment effect is approximately:

$$Var(\hat{\eta}_{0010}) = \frac{\tau_{\gamma 111} + 4\left\{\left[\tau_\beta + (\tau_\pi + \sigma^2 / n)/J\right]/K\right\}}{L}. \quad [10.6]$$

If the data are balanced, we can use the results of a nested analysis of variance with random effects for the classrooms, schools, and sites, and fixed effects for the treatment. The test statistic is an F statistic. When the null hypothesis is false, the F statistic follows the non-central F distribution, F(1, L-2; λ). The non-centrality parameter is shown below:

$$\lambda = \frac{L\eta_{0010}^2}{\tau_{\gamma 111} + 4\left\{\left[\tau_\beta + (\tau_\pi + \sigma^2 / n)/J\right]/K\right\}}. \quad [10.7]$$

Recall that the greater the non-centrality parameter, the greater the power of the test. From equation 6.7, it is clear that the number of sites has the greatest affect on the non-centrality parameter followed by the number of schools, classrooms, and finally students.

10.3Standardized notation

In the standardized model, the between-site variance is removed and the sum of the level-one, level-two, and level-three variances is set to 1. Similar to the three-level cluster randomized trial, there are two ICC's, ρ_{level2} and ρ_{level3}. The first ICC, ρ_{level2}, corresponds to the between-classroom variance relative to the between and within school variance, $\rho_{level2} = \frac{\tau_\pi}{\tau_\pi + \tau_\beta + \sigma^2}$.

The second ICC, ρ_{level3}, corresponds to the between-school variance relative to the between and within school variance, $\rho_{level3} = \frac{\tau_\beta}{\tau_\pi + \tau_\beta + \sigma^2}$. The standardized main effect of treatment, δ, is

$\delta = \frac{\eta_{0010}}{\sqrt{\tau_\pi + \tau_\beta + \sigma^2}}$, and the variance of the standardized treatment effect, σ_δ^2, is

$\sigma_\delta^2 = \frac{\tau_{\gamma 111}}{\tau_\pi + \tau_\beta + \sigma^2}$. The non-centrality parameter, λ, can be rewritten as:

$$\lambda = \frac{L\delta_{0010}^2}{\sigma_\delta^2 + 4\left\{\rho_{level3} + \left[\rho_{level2} + (1 - \rho_{level3} - \rho_{level2}/n)/J\right]/K\right\}}.^6 \quad [10.8]$$

10.4The model with a level-3 covariate

[6] The Optimal Design asks for slightly different parameters than those presented in the non-centrality parameter. The Optimal Design asks the user to enter the variance components prior to blocking and the percent of variance explained by blocking and calculates the parameters for the noncentrality parameter within the program.

The level-1 and level-2 models remain the same when we add a level-3 covariate. The level-3 model with a school-level covariate is:

$$\beta_{00kl} = \gamma_{000l} + \gamma_{001l} W_{00kl} + \gamma_{002l} S_{00kl} + u_{00kl} \qquad u_{00kl} \sim N(0, \tau_{\beta_{|S}}) \qquad [10.9]$$

$$\tau_{\beta|s} = (1 - \rho^2_{s\beta_{00k}})\tau_\beta$$

where γ_{000l} is the mean for school l;

γ_{001l} is the treatment effect at school l;

γ_{002l} is the regression coefficient for the school-level covariate, which is assumed constant across sites;

W_{00kl} is the indicator variable where -1/2 represents the control group and ½ represents the treatment group;

u_{00kl} is the random effect associated with school;

$\rho_{s\beta_{00k}}$ is the correlation between the school-level covariate and the school-level mean;

τ_β is the variance between schools within sites; and

$\tau_{\beta|S}$ is the variance between schools within sites conditional on the school-level covariate.

The level-4 model is:

$$\begin{aligned} \gamma_{000l} &= \eta_{0000} + s_{000l} \\ \gamma_{001l} &= \eta_{0010} + s_{001l} \\ \gamma_{002l} &= \eta_{0020} \end{aligned} \qquad \begin{aligned} \text{var}(s_{000l}) &\sim N(0, \tau_{\gamma_{000}}) \\ \text{var}(s_{001l}) &\sim N(0, \tau_{\gamma_{011}}) \end{aligned} \qquad [10.10]$$

where η_{0000} is the grand mean;

η_{0010} is the average treatment effect;

η_{0020} is the average effect of the regression coefficient;

s_{000} is the random effect associated with each site mean;

s_{001} is the random effect associated with each site treatment effect;

$\tau_{\gamma 000}$ is the variance between site means;

$\tau_{\gamma 111}$ is the variance between sites on the treatment effect; and

$\tau_{\gamma 001}$ is the covariance between site-specific means and site-specific treatment effects.

10.5Testing the treatment effect (including a level-3 covariate)

The average treatment effect is estimated by:

$$\hat{\eta}_{0010} = \bar{Y}_E - \bar{Y}_C - \hat{\eta}_{0020}(\bar{S}_E - \bar{S}_C) \qquad [10.11]$$

where $\bar{Y}_E$ is the mean for the experimental group;

$\bar{Y}_C$ is the mean for the control group;

$\bar{S}_E$ is the covariate mean for the experimental group; and

$\bar{S}_C$ is the covariate mean for the control group.

If the data are balanced, we can use the results of a nested analysis of variance with random effects for the classrooms, schools, and sites, and fixed effects for the treatment. The test statistic is an F statistic. When the null hypothesis is false, the F statistic follows the non-central F distribution, F(1, *L*-1; λ_s). The non-centrality parameter is shown below:

$$\lambda_s = \frac{L\eta_{0010}^2}{\tau_{\gamma 111} + 4(\tau_{\beta|s} + (\tau_\pi + \sigma^2/n)/J)/K)}. \qquad [10.12]$$

We can rewrite 10.12 in standardized notation as follows:

$$\lambda_s = \frac{L\delta_{0010}^2}{\sigma_\delta^{*2} + 4\left\{\rho^*_{level\,3} + \left[\rho^*_{level\,2} + (1 - \rho^*_{level\,3} - \rho^*_{level\,2}/n)/J\right]/K\right\}} \qquad [10.13]$$

where $\rho^*_{level\,2} = \dfrac{\tau_\pi}{\tau_\pi + \tau_{\beta|s} + \sigma^2}$ is the conditional level two ICC;

$\rho^*_{level\,3} = \dfrac{\tau_{\beta|s}}{\tau_\pi + \tau_{\beta|s} + \sigma^2}$ is the conditional level three ICC;

$\delta^* = \dfrac{\eta_{0010}}{\sqrt{\tau_\pi + \tau_{\beta|s} + \sigma^2}}$ is the standardized main effect of treatment conditional on the covariate; and

$\sigma_\delta^{*2} = \dfrac{\tau_{\gamma 111}}{\tau_\pi + \tau_{\beta|s} + \sigma^2}$ is the variance of the standardized treatment effect conditional on the covariate.

10.6 The fixed effects model

The fixed effects model is the same as the random effect model except that the site-specific effects are fixed constants instead of random variables. The new level-4 model is:

$$\begin{aligned}\gamma_{000l} &= \eta_{0000} + s_{000l} \\ \gamma_{001l} &= \eta_{0010} + s_{001l} \\ \gamma_{002l} &= \eta_{0020}\end{aligned} \qquad [10.14]$$

where η_{0000} is the grand mean;

η_{0010} is the average treatment effect;

η_{0020} is the average effect of the regression coefficient;

s_{000l}, are fixed effects associated with each site mean, constrained to have a mean of zero; and

s_{001l} are fixed effects associated with each site treatment effect, constrained to have a mean of zero.

Similar to the three-level multi-site cluster randomized trial, the main effect of treatment, η_{0010}, and the fixed treatment-by-site interaction effects s_{001l}, are of primary interest.

10.7 Testing the average treatment effect

If the data are balanced, we can use the results of a nested analysis of variance with random effects for the classrooms and schools and fixed effects for sites, treatments, and site-by-treatment interaction. Similar to prior tests, the test statistic is an *F* statistic. The *F* test follows a non-central *F* distribution, $F(1, L(K\text{-}2); \lambda)$. Recall that the non-centrality parameter is a ratio of the squared-treatment effect to the variance of the treatment effect estimate. Below is the non-centrality parameter for the test in standardized notation without a school-level covariate:

$$\lambda = \frac{L\delta_{0010}^2}{4\left\{\rho_{level3} + \left[\rho_{level2} + (1 - \rho_{level3} - \rho_{level2}/n)/J\right]/K\right\}} \text{ (see footnote 6)} \qquad [10.15]$$

where $\rho_{level2} = \dfrac{\tau_\pi}{\tau_\pi + \tau_\beta + \sigma^2}$ is the variance between clusters relative to the between and within school variation within;

$\rho_{level3} = \frac{\tau_\beta}{\tau_\pi + \tau_\beta + \sigma^2}$ is the variance between schools relative to the between and within school variation within blocks; and

$\delta = \frac{\eta_{0010}}{\sqrt{\tau_\pi + \tau_\beta + \sigma^2}}$ is the standardized main effect of treatment.

Adjusting for the covariate reveals:

$$\lambda_s = \frac{L\delta^{*2}}{4\left\{\rho^*_{level3} + \left[\rho^*_{level2} + (1 - \rho^*_{level3} - \rho^*_{level2}/n)/J\right]/K\right\}} \qquad [10.16]$$

where $\rho^*_{level2} = \frac{\tau_\pi}{\tau_\pi + \tau_{\beta_{|s}} + \sigma^2}$ is the variance between clusters relative to the between and within school variation within blocks conditional on the school-level covariate, S;

$\rho^*_{level3} = \frac{\tau_{\beta|s}}{\tau_\pi + \tau_{\beta|s} + \sigma^2}$ is the variance between schools relative to the between and within school variation within blocks conditional on the school-level covariate, S; and

$\delta^* = \frac{\eta_{0010}}{\sqrt{\tau_\pi + \tau_{\beta|s} + \sigma^2}}$ is the standardized main effect of treatment conditional on the school-level covariate, S.

Note that unlike the case of the random effects model, σ^2_δ, the variance of the treatment effect, does not appear in the denominator of the non-centrality parameter. However, as in the case of the random effects model, if the variation of the treatment effects across sites is large, the average treatment effect may not be informative.

10.8 Using Optimal Design for multisite cluster randomized trials (4 level model)

The menu for the MSCRT with treatment at level-3 is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 3.

Power for treatment effect on y-axis

Power vs. cluster size (*n*)

Power vs. number of clusters per school (*J*)

Power vs. number of schools per site (*K*)

Power vs. total number of sites (*L*)

Power vs. effect size (delta)

Power vs. proportion of explained variation by level-two covariate (R2)

MDES fr treatment effect on y-axis

MDES vs. cluster size (n)

MDES vs. number of clusters per school (J)

MDES vs. number of schools per site (K)

MDES vs. total number of sites (L)

MDES vs. power (P)

MDES vs. proportion of explained variation by level-two covariate (R2)

The first set of options present the power on the y-axis and the second set of options presents the MDES on the y-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

10.9 Example

Suppose a team of researchers develop a new literacy program. The founders of the new program propose that students who participate in the program will have increased reading achievement. They propose a four-level design with students nested within classrooms within schools within districts and they want to block by district. That is, within each district, half of the schools will be randomly assigned to the new program and half to the current program. They plan to test students who are in classrooms that participate in the new program (experimental group) and students who are in classrooms that participate in the regular program (control group) in each of the schools using a reading test to determine if students using the new program score higher.

10.10 Power determination approach for conducting a power analysis

Based on past studies, the researchers expect about 20 percent of the variation in the outcome to be between schools and 10 percent to be between classrooms within schools (prior to blocking). By blocking on district, the researchers expect to explain 40% of the variation in the outcome variable. They are interested in detecting an effect size of at least 0.25 with adequate power. Assuming that they plan to test 20 students per classrooms and have secured 8 classrooms per school and 8 schools per district, how many districts (sites) are necessary to achieve power = 0.80? Assume the researchers are interested in generalizing to a broader population of districts thus they treat the districts as random effects and assume an effect size variability of 0.01.

Suppose the researchers have access to a school level pre-test that explain about 49% of the variation in post-test scores. How many districts are required after including the school-level covariate?

In Scenario 1, the number of sites, *L*, is unknown. As a result, we want to select the power vs. number of sites (*L*) option. This allows the number of sites to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 3 → Power on for treatment effect on y-axis → Power vs. number of sites (*L*) as shown in Figure 10.1.

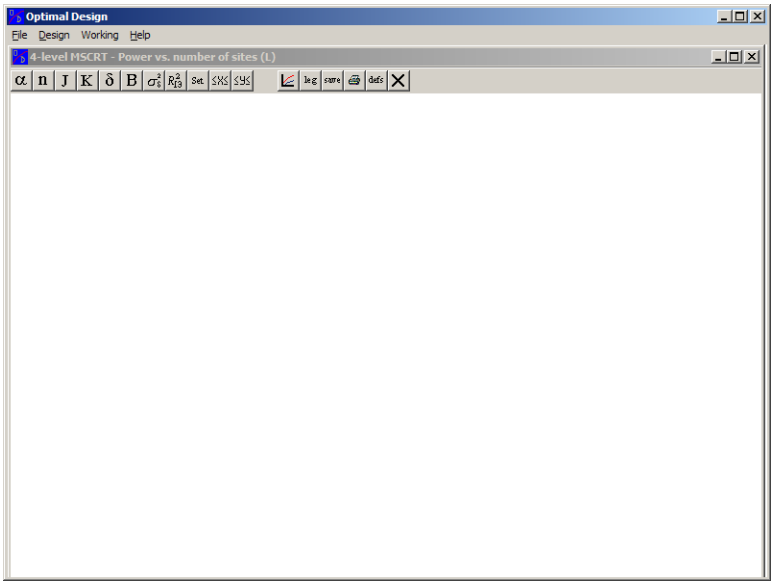


Figure 10.1. Initial screen for power vs. number of sites (*L*).

The toolbar at the top includes the parameters required for calculating the power. The number of sites (*L*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 20. By clicking on n(1) = 20, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J*(1) = 8. This is the number of classrooms per school.

Step 4: Click on *K*. Set *K*(1) = 8. The is the number of schools per district.

Step 5: Click on δ. Set delta (1) = 0.25.

Step 6: Click on B. Set B(1) = 0.40. This defines the percent of variance explained by blocking on district.

Step 7: Click on σ_δ^2. Set sigma2d = 0.01. Setting the effect size variability greater than 0 assumes random site effects.

Step 6: Click on set. Set $\tau_\pi = 0.10$ and $\tau_\beta = 0.20$. This sets the variance components at level 2 and 3.

The resulting power curve appears in Figure 10.2.

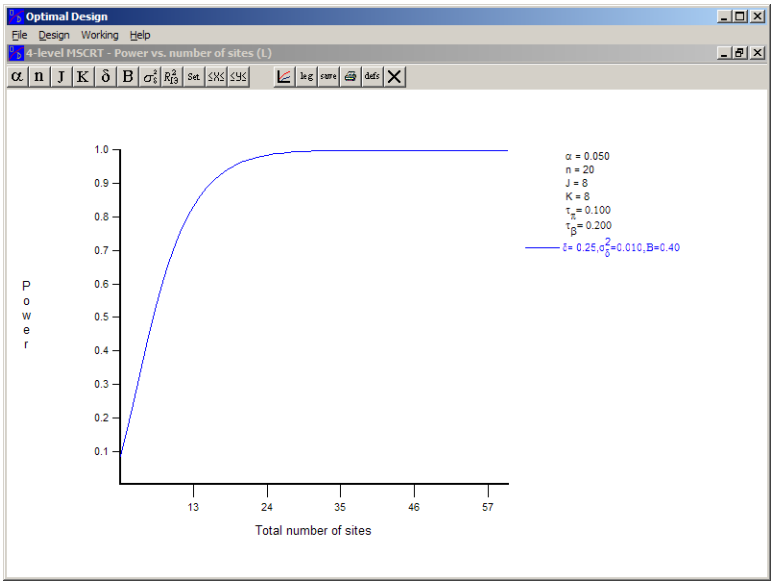


Figure 10.2. Power curve.

Clicking along the power curve, we can see that 13 sites or districts are required for the study. Within each district, we randomly assign 4 schools to the treatment condition and 4 schools to the control condition. Let's see what happens when we include the school-level covariate.

Step 8: Click on R. Set R2 – 2 equal to 0.49. Two power curves appear as shown in Figure 10.3.

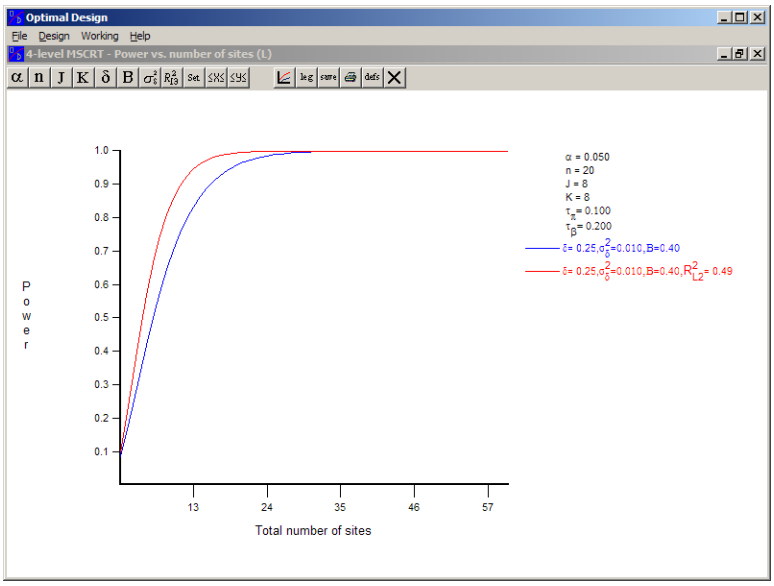


Figure 10.3. Two power curves.

According to the key, the dotted trajectory is the power curve when the covariate is included. In this case, approximately 8 districts are needed, a reduction of 4 districts and 40 schools, which may greatly reduce the cost of the study.

If we were not interested in generalizing to a larger population of districts, we could consider the districts as fixed effects. In this case, we would set $\sigma_\delta^2 = 0$ as shown in Figure 10.4.

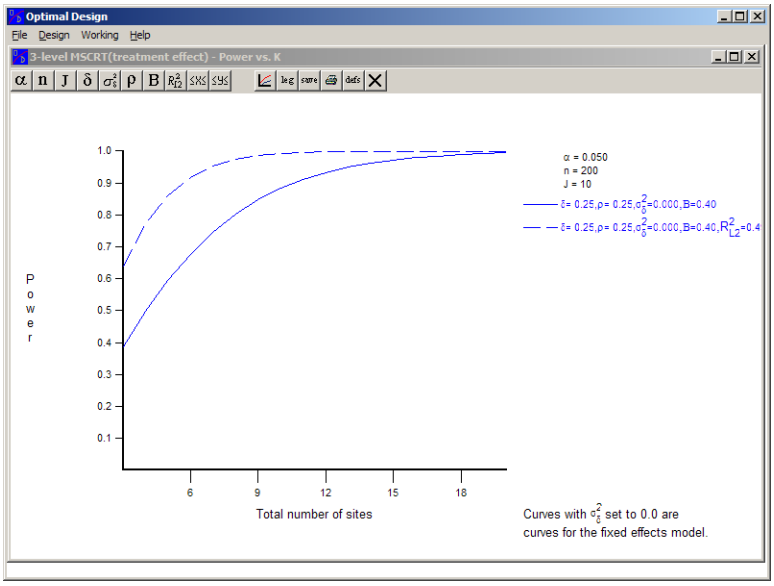


Figure 10.4. Power curve with fixed effects.

Clicking on the curve reveals that 10 sites, with and without a covariate, are required for power = 0.80.

The example provided in this section placed the number of districts on the x-axis. However, any of the other parameters could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

10.11 Effect size approach for conducting a power analysis

Based on past studies, the researchers expect about 20 percent of the variation in the outcome to be between schools and 10 percent to be between classrooms within schools (prior to blocking). By blocking on district, the researchers expect to explain 40% of the variation in the outcome variable. They are interested in detecting an effect size of at least 0.25 with adequate power. Assuming that they plan to test 20 students per classrooms and have secured 8 classrooms per school, 8 schools per district, and 10 districts, what is the MDES for power = 0.80? Assume the researchers are interested in generalizing to a broader population of districts thus they treat the districts as random effects and assume an effect size variability of 0.01. Suppose the researchers have access to a school level pre-test that explain about 49% of the variation in post-test scores. What is the MDES after including the covariate?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. total number of sites (L). This allows the user to see how the MDES changes as a function of the power holding all other parameters constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Multi-site (or blocked) cluster randomized trials → Treatment at level 3 → MDES for treatment effect on y-axis → MDES vs. total number of sites (L) as shown in Figure 10.5.

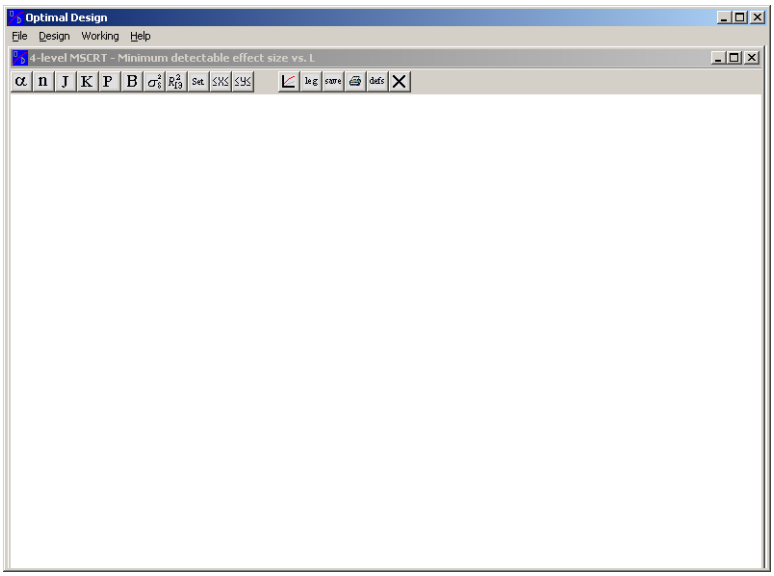


Figure 10.5. Initial screen for multi-site cluster randomized trials.

Step 2: Click on n. Set n(1) = 200. By clicking on n(1) = 200, the default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J*(1) = 8. This is the number of classrooms per school.

Step 4: Click on *K*. Set *K*(1) = 8. This is the number of schools per district.

Step 4: Click on *P*. Set *P* = 0.80.

Step 5: Click on B. Set B(1) = 0.40. This defines the percent of variance explained by blocking on district.

Step 6: Click on σ_δ^2. Set sigma^2 = 0.01. Setting the effect size variability greater than 0 assumes random site effects.

Step 7: Click on set. Set $\tau_\pi = 0.10$ and $\tau_\beta = 0.20$. This sets the variance components at level 2 and 3. The resulting curve is in Figure 10.6.

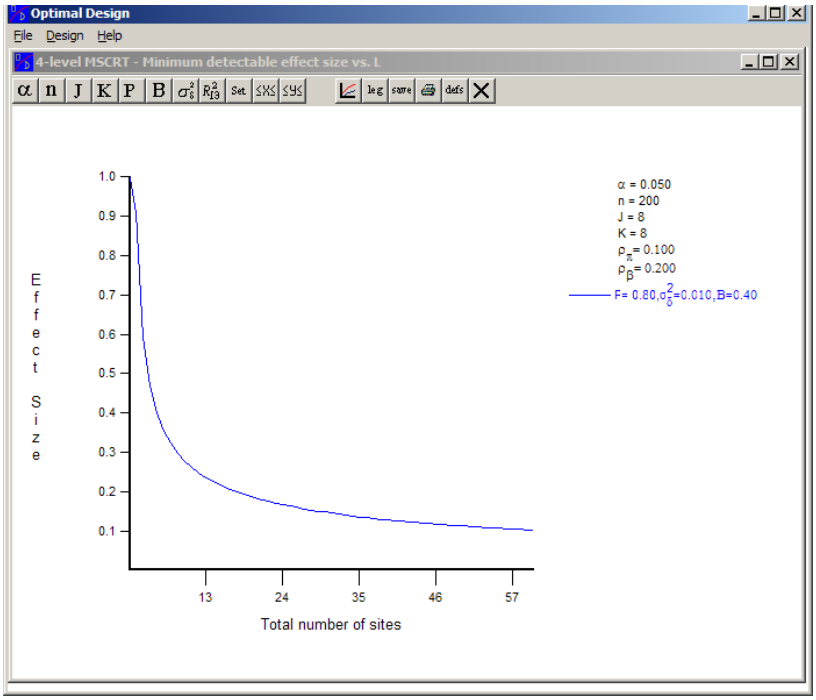


Figure 10.6. Power curve.

Clicking along the trajectory we can see that for 10 districts, the MDES is approximately 0.28.

Let's see what happens when we add a cluster-level covariate.

Step 8: Click on R. Set R – 2 = 0.49. Figure 10.7 displays the two power curves.

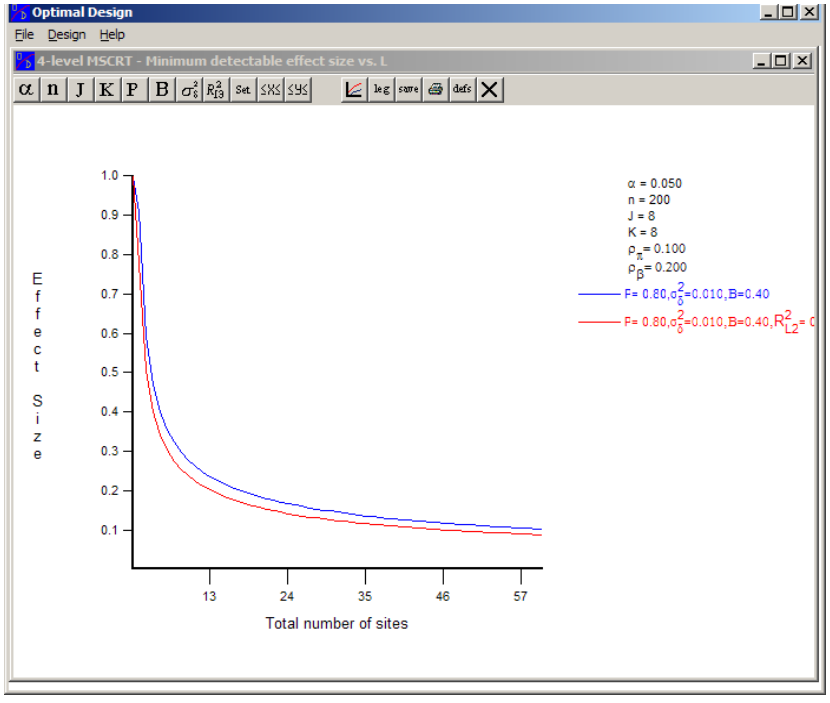


Figure 10.7. Power curve with cluster-level covariate.

Clicking along the dotted trajectory, the MDES with power = 0.80 is 0.23, reducing the MDES by 0.05 effect size units.

Another option is to treat the sites as fixed effects. By clicking on σ_δ^2 and setting it to 0, the sites are assumed to be fixed effects. Figure 10.8 displays the power for this case.

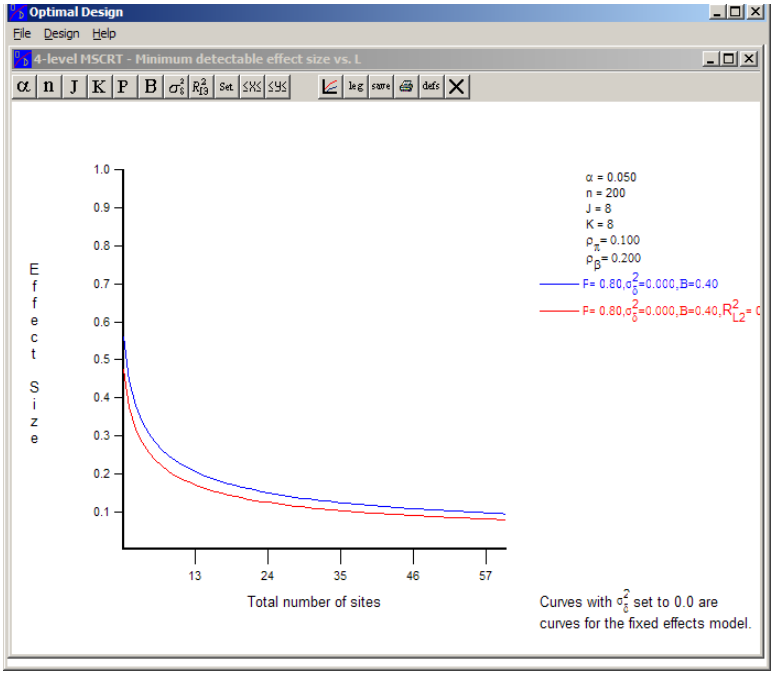


Figure 10.8. Power curve with fixed effects.

The MDES reduces to 0.24 and 0.19, without and with the covariate. However, the generalizability also changes which is not a statistical issue but very important in terms of the study design.

11.0 Cluster randomized trials with repeated measures

In a typical longitudinal study, an observation is recorded prior to treatment, often referred to as the baseline measurement, and then after the treatment a pre-determined number of times. Measuring participants prior to treatment and post-treatment allows the researchers to assess individual growth. Individual growth may be plotted via a straight line or a curvilinear trajectory. A linear trajectory, or first degree polynomial, is characterized by an intercept and a linear rate of change, or slope. Curvilinear trajectories are second, third, or higher degree polynomials. A second degree polynomial, also known as a quadratic polynomial, adds an acceleration parameter to the intercept and rate of change. A third degree polynomial, also known as a cubic polynomial, is characterized by 4 parameters, change in acceleration, rate of acceleration, linear rate of change, and an intercept.

In a simple repeated measures design, individuals are repeatedly observed and individual trajectories are plotted to assess average treatment effects on a specific polynomial change parameter. In this chapter we extend the simple design to settings in which individuals are nested within clusters and treatment is applied at the cluster level. This allows us to assess the average difference in the polynomial change parameter for those in the treatment group and those in the control group, accounting for the cluster effect.

The power to detect the main effect of treatment in a repeated measure cluster randomized trial is more complicated than in a cluster randomized trial because we need to take the repeated measures on each person into consideration. For simplicity, we assume orthogonal designs with continuous outcomes, a random-effects covariance structure, homogeneous covariance structures within treatments, and complete data. The data lend to the three-level hierarchical model described in the next section.

11.1 The model

Data from a cluster randomized trial with repeated measures on the individuals can be represented with a three-level model, with occasions nested within persons and persons nested within clusters. The general level-1 model, or repeated measures model, represents the trajectory of change for person i as a polynomial function of degree $P-1$ defined at equally spaced observations. The model is:

$$Y_{mij} = \sum_{p=0}^{P-1} \pi_{pij} c_{pm} + e_{mij}, \qquad e_{mij} \sim N(0, \sigma^2) \qquad [11.1]$$

for $m \in \{1,2,..., M\}$ observations, $i \in \{1,2,..., n\}$ persons, and $j \in \{1,2,..., J\}$ clusters, where

p is the polynomial order of change (e.g., linear, quadratic, or cubic);

π_{pij} is the level-1 coefficient for the polynomial of order p;

c_{pm} is the orthogonal polynomial contrast coefficient;

e_{mij} is the error associated with the repeated measures; and

σ^2 is the within-person variance.

Note the orthogonal polynomial contrast coefficients are necessary to center the data. These coefficients are given by (see, e.g., Kirk 1982; Raudenbush and Liu 2001):

$$c_{0m} = 1, \qquad [11.2]$$

$$c_{1m} = m - \sum_{m=1}^{M} m / M,$$

$$c_{2m} = \frac{1}{2}\left(c_{1m}^2 - \sum_{m=1}^{M} c_{1m}^2 / M \right), \text{ and}$$

$$c_{3m} = \frac{1}{6}\left(c_{1m}^3 - \frac{\sum_{m=1}^{M} c_{1m}^4}{\sum_{m=1}^{M} c_{1m}^2} c_{1m} \right).$$

The level-2 model, or person-level model, is:

$$\pi_{pij} = \beta_{p0j} + r_{pij}, \qquad r_{pij} \sim N(0, \tau_{\pi p}) \qquad [11.3]$$

where β_{p0j} is the cluster mean for the p^{th} polynomial change parameter;

r_{pij} is the random effect associated with the persons; and

$\tau_{\pi p}$ is the between-person variance for the p^{th} polynomial change parameter.

The level-3 model, or cluster level model is:

$$\beta_{p0j} = \gamma_{p00} + \gamma_{p01} W_j + u_{p0j}, \qquad u_{p0j} \sim N(0, \tau_{\beta p}) \qquad [11.4]$$

where γ_{p00} is the grand mean for the polynomial order of change;

γ_{p01} is the main effect of treatment;

W_j is a treatment contrast indicator, ½ for treatment and -½ for control;

u_{p0j} is the random effect associated with each cluster; and

$\tau_{\beta p}$ is the between-cluster variance for the polynomial order of change.

To help clarify the general model, consider a first degree polynomial order of change, or linear model $(p=1)$. The level-1 model is:

$$Y_{mij} = \pi_{0ij} + \pi_{1ij} c_{1m} + e_{mij}, \qquad e_{mij} \sim N(0, \sigma^2) \qquad [11.5]$$

for $m \in \{1,2,...,M\}$ occasions, $i \in \{1,2,...,n\}$ persons, and $j \in \{1,2,...,J\}$ clusters,

where π_{0ij} is the mean response for person *i* in cluster *j* on occasion *m*;

π_{1ij} is the average rate of change for person *i* in cluster *j* on occasion *m*;

c_{1m} is the orthogonal linear contrast coefficient;

e_{mij} is the error associated with the repeated measures; and

σ^2 is the within-person variance.

Note that in the case of the linear model, the contrast coefficients are easily computed using the formulas in equation 2. For example, if *M*=5, the orthogonal contrast coefficients for a first degree polynomial are:

$$\begin{aligned} c_0 &= (1,1,1,1,1) \\ c_1 &= (-2,-1,0,1,2) \end{aligned} \qquad [11.6]$$

The level-2 model, or person level model is:

$$\begin{aligned} \pi_{0ij} &= \beta_{00j} + r_{0ij} \qquad r_{0ij} \sim N(0, \tau_{\pi 0}) \\ \pi_{1ij} &= \beta_{10j} + r_{1ij} \qquad r_{1ij} \sim N(0, \tau_{\pi 1}) \end{aligned} \qquad [11.7]$$

where β_{00j} is the mean response in cluster *j*;

β_{10j} is the average growth rate in cluster *j*;

r_{0ij} is the random effect associated with the mean response for person *i* in cluster *j*;

r_{1ij} is the random effect associated with the growth rate for person *i* in cluster *j*;

$\tau_{\pi 0}$ is the between-person variance in means; and

$\tau_{\pi 1}$ is the between-person variance in growth rates.

The level-3 model, or cluster level model is:

$$\beta_{00j} = \gamma_{000} + \gamma_{001} W_j + u_{00j} \qquad u_{00j} \sim N(0, \tau_{\beta 0}) \qquad [11.8]$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101} W_j + u_{10j} \qquad u_{10j} \sim N(0, \tau_{\beta 1})$$

where γ_{000} is the grand mean;

γ_{001} is the main effect of treatment for the mean;

W_j is the treatment indicator, ½ for treatment and -½ for control;

γ_{100} is the average growth rate;

γ_{101} is the main effect of treatment for the growth rates;

u_{00j} is the random effect associated with the mean for each cluster;

u_{10j} is the random effect associated with the growth rate for each cluster;

$\tau_{\beta 0}$ is the between-cluster variance in means; and

$\tau_{\beta 1}$ is the between-cluster variance in growth rates.

Note that for a first degree polynomial, our primary interest is in growth rates, thus we are interested in γ_{101}, the main effect of treatment on the growth rates, and in $\tau_{\beta 1}$, the between-cluster variance in growth rates.

11.2 Testing the treatment effect

The average treatment effect for the p^{th} polynomial order of change in our balanced design is defined at level 3. It is estimated by:

$$\hat{\gamma}_{p01} = \frac{\sum_{j \in E} \sum_{i=1}^{n} \hat{\pi}_{1ij}}{nJ_E} - \frac{\sum_{j \in C} \sum_{i=1}^{n} \hat{\pi}_{1ij}}{nJ_C}. \qquad [11.9]$$

where $\hat{\pi}_{1ij} = \dfrac{\sum_{m=1}^{M} c_{1m} Y_{mij}}{\sum_{m=1}^{M} c_{1m}^2}$

is the ordinary least squares estimate of the person-specific linear growth rate π_{1ij}.

The variance of the treatment effect for the p^{th} polynomial order of change (Raudenbush & Liu, 2001) is:

$$Var(\hat{\gamma}_{p01}) = \frac{4[\tau_{\beta p} + (\tau_{\pi p} + V_p)/n]}{J}, \qquad [11.10]$$

$$V_p = \frac{\sigma^2}{\sum_{m=1}^{M} c_{pm}^2} = \frac{\sigma^2 f^{2p}(M-p-1)!}{K_p(M+p)!},$$

where f is the frequency of observation;

D is the duration of the study;

M is the total number of occasions, $M=Df+1$;

p is the polynomial order of change; and

K_p is a constant, where K_1=1/12, K_2=1/720, and K_3=1/100,800.

Note that V_p denotes the conditional variance of the least squares estimate of each participant's change parameter.

We can translate the above formulas to a more concrete example in the case of a first degree polynomial. For a first degree polynomial, the variance of the estimate of the treatment effect is:

$$Var(\hat{\gamma}_{101}) = \frac{4[\tau_{\beta 1} + (\tau_{\pi 1} + V_1)/n]}{J} \qquad [11.11]$$

where $V_1 = \frac{\sigma^2}{\sum_{m=1}^{M} c_{1m}^2} = \frac{\sigma^2 f^2 (M-2)!}{1/12(M+1)!}$. [11.12]

In the general case, we can use the following hypotheses to test the significance of the main effect of treatment for the polynomial order of interest:

$$H_0 : \gamma_{p01} = 0$$
$$H_1 : \gamma_{p01} \neq 0 \qquad [11.13]$$

When the null hypothesis is true, the test statistic is an F statistic and follows a central F distribution, $F(1, J\text{-}2)$. The test statistic is:

$$F = \frac{\hat{\gamma}_{p01}}{Var(\hat{\gamma}_{p01})}. \qquad [11.14]$$

When the alternative hypothesis is true, the test statistic remains the same but follows a noncentral *F* distribution, $F(1, J\text{-}2; \lambda)$. Recall that the noncentrality parameter is a ratio of the squared treatment effect to the variance of the treatment effect estimate. The noncentrality parameter is:

$$\lambda = \frac{\gamma_{p01}^2}{Var(\hat{\gamma}_{p01})} = \frac{J\gamma_{p01}^2}{4[\tau_{\beta p} + (\tau_{\pi p} + V_p)/n]} \qquad [11.15]$$

Recall that the larger the noncentrality parameter, the greater the power of the test. Looking at the formula, we can see that *J* is the most influential sample size for increasing the power. In other words, the number of clusters in more important than the number of people within each cluster to increase the power. It is particularly important to have a large number of clusters if there is a lot of between-cluster variation, $\tau_{\beta p}$. Also, increasing the number of occasions, *M*, reduces the within-person variance, which increases the power. Note that *M* is a function of *f* and *D*, where $M=(fD+1)$ so increasing the frequency of the observations or duration of the study increases *M*. Increasing *n*, the number of people within each cluster, will also decrease the total within and between-person variance, thus increasing the power. Finally, larger effect sizes increase the power to detect a treatment effect.

11.3 Standardized notation

Thus far, we have concentrated on the unstandardized model. However, similar to cluster randomized trials, and without loss of generality, researchers typically use standardized notation. Let's see how we translate the parameters to standardized notation.

The standardized effect size for a polynomial of order *p* is:

$$\delta = \frac{\gamma_{p01}}{\sqrt{\tau_{\beta p} + \tau_{\pi p}}} \qquad [11.16]$$

where

γ_{p01} is the main effect for the polynomial order of change, and

$\tau_{\beta p} + \tau_{\pi p}$ is the total between-cluster and between-person variance, denoted τ.

In words, δ is the group difference on the polynomial of interest divided by the standard deviation for that polynomial, or the square root of the sum of the between-cluster variance and the between-person variance for the specified polynomial. Similar to standardized models we

defined in previous chapters, we need to define ρ, the intra-class correlation. The intra-class correlation, ρ, is:

$$\rho = \frac{\tau_{\beta p}}{\tau_{\beta P} + \tau_{\pi p}} \qquad [11.17]$$

where $\tau = \tau_{\beta p} + \tau_{\pi p}$ is the total between-cluster and within-cluster variance;

$\tau_{\beta p}$ is the between-cluster variance on the polynomial of interest; and

$\tau_{\pi p}$ is the within-cluster variance on the polynomial of interest.

Note that if $\tau = 1$, then $\tau_{\beta p} = \rho$ and $\tau_{\pi p} = 1 - \rho$ which is consistent with the intra-class correlation for a cluster randomized trial. Also, ρ is a ratio of the between-cluster variance to the total variance for a specific polynomial order of change. We can think of ρ as partitioning the growth-rate variance into a between-cluster and within-cluster component.

Using the standardized effect size, δ, ρ, and constraining $\tau = 1$, we can rewrite the variance of the treatment effect estimate as

$$Var(\hat{\gamma}_{p01}) = \frac{4[\rho + (1 - \rho + V_p)/n]}{J}. \qquad [11.18]$$

Another simplification involves rewriting the variance in terms of the reliability of the person-specific polynomial change. The reliability is denoted α_p and is defined as:

$$\alpha_p = \frac{\tau_{\pi p}}{\tau_{\pi p} + V_p}. \qquad [11.19]$$

Rewriting the variance in terms of the reliability we get:

$$Var(\hat{\gamma}_{p01}) = \frac{4[\rho + (1 - \rho)/(\alpha_p n)]}{J}. \qquad [11.20]$$

We write the variance in this form because standard programs for hierarchical data often give us an estimate of the person specific reliability.

We can also rewrite the noncentrality parameter using standardized notation. The new noncentrality parameter is:

$$\lambda = \frac{J\delta^2}{4[\rho + (1 - \rho)/(\alpha_p n)]} \quad . \qquad [11.21]$$

Note that the power is now a function of the number of clusters, *J*, the cluster size, *n*, the standardized effect size, δ, the intra-class correlation, ρ, the reliability, α_p, which is a function of the between-person variance, $\tau_{\pi p}$, the within-person variance, σ^2, the study duration, *D*, the frequency of the observations, *f*, and the number of occasions, *M*.

11.4 Using Optimal Design for cluster randomized trials with repeated measures

The cluster randomized trial with repeated measures option allows the researcher to explore the power for the main effect of treatment as a function of the cluster size, *n*, the number of clusters, *J*, the intra-class correlation, ρ, and the desired effect size, δ. Below is the menu.

Power on y-axis

Power vs. Cluster Size (*n*)

Power vs. Number of Clusters (*J*)

Power vs. Intra-class Correlation (ρ)

Power vs. Effect Size (δ)

MDES on y-axis

MDES vs. Cluster Size (*n*)

MDES vs. Number of Clusters (*J*)

MDES vs. Intra-class Correlation (ρ)

MDES vs. Power (*P*)

11.5 Example

Imagine that a group of researchers develop a new phonics program for first graders. The program is an intense year-long program. The researchers propose a repeated measures design for students nested within schools. They plan to assess students at the beginning of the year, prior to treatment, and then on six occasions throughout the year. Researchers are interested in the growth rate of students so they propose a linear model. Past research suggests that 15 percent of the total variation in the outcome is between schools. The researchers conducted a pilot study and found estimates of the within person variability to be 1.0 and the overall variability in growth rates to be 1.0. Section 11.6 presents a scenario in which the power determination approach to conducting a power analysis is the most appropriate and provides the details of how to do the power analysis. Section 11.7 presents a scenario in which the effect size approach for

conducting a power analysis is most appropriate and provides the details of how to do the power analysis.

11.6 Power determination approach for conducting a power analysis

In scenario 1, the researchers are interested in detecting an effect size of 0.30. They plan to assess 20 students in each school and estimate an intraclass correlation of 0.15. How many schools are necessary for power = 0.80?

In Scenario 1, the total number of clusters (*J*) is unknown so we select the power vs. total number of clusters (*J*) option. This allows the number of clusters to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trial with person level outcomes → Cluster randomized trials → Repeated measures → Power on the y-axis → Power vs. total number of clusters (*J*). Figure 11.1displays the screen.

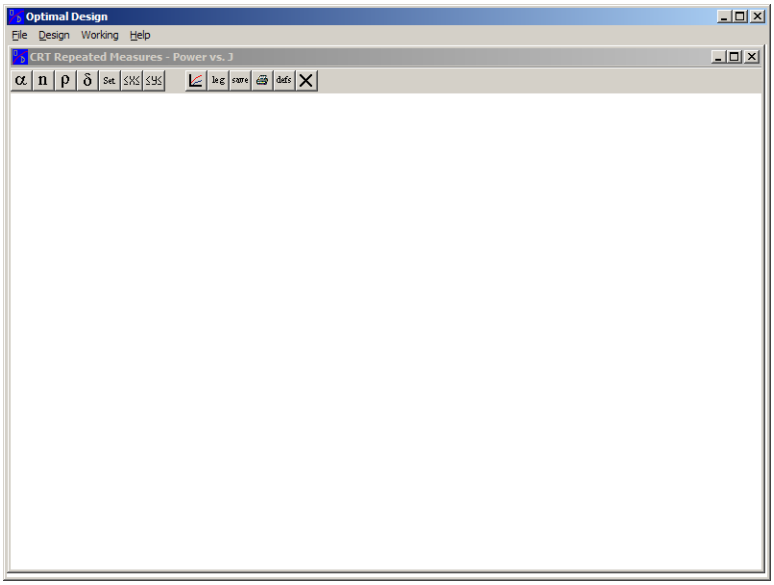


Figure 11.1. Initial screen for cluster randomized trials with repeated measures.

Step 2: Click on n. Set n(1) = 20.

Step 3: Click on ρ. Set ρ (1) = 0.15.

Step 4: Click on δ. Set δ (1) = 0.30.

Step 5: Click on set. Set f = 1, d = 6, variability of level-1 residual = 1, variability of level-1 coefficient = 1, select polynomial order linear. The final curve appears in Figure 11.2.

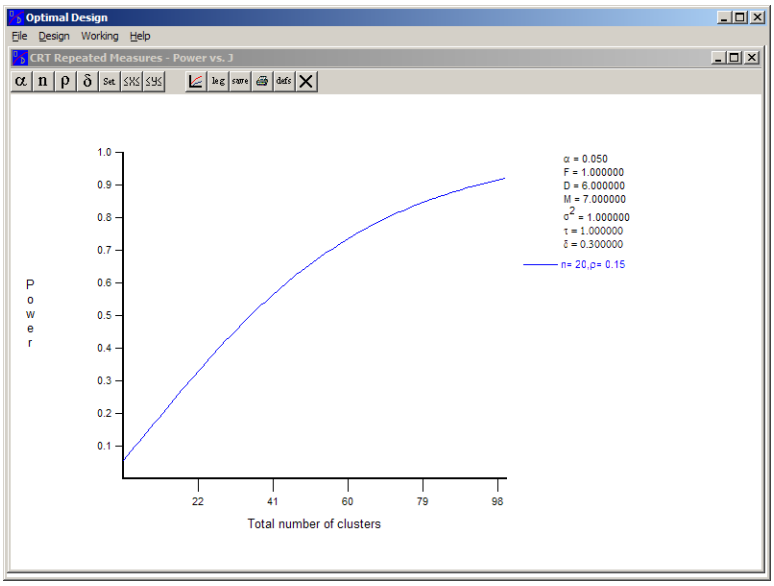


Figure 11.2. Power curve.

Clicking along the trajectory reveals that approximately 72 clusters are required for the study, that is, 36 in the treatment condition and 36 in the control condition.

11.6 Effect size approach for conducting a power analysis

In scenario 2, the researchers are have secured 60 schools and they plan to assess 20 students in each school. They estimate an intraclass correlation of 0.15. What is the MDES for power = 0.80?

In Scenario 2, the MDES is unknown so we select the MDES vs. total number of clusters (*J*) option. This allows the power to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trial with person level outcomes → Cluster randomized trials → Repeated measures → MDES on the y-axis → MDES vs. total number of clusters (*J*). Figure 11.3 displays the screen.

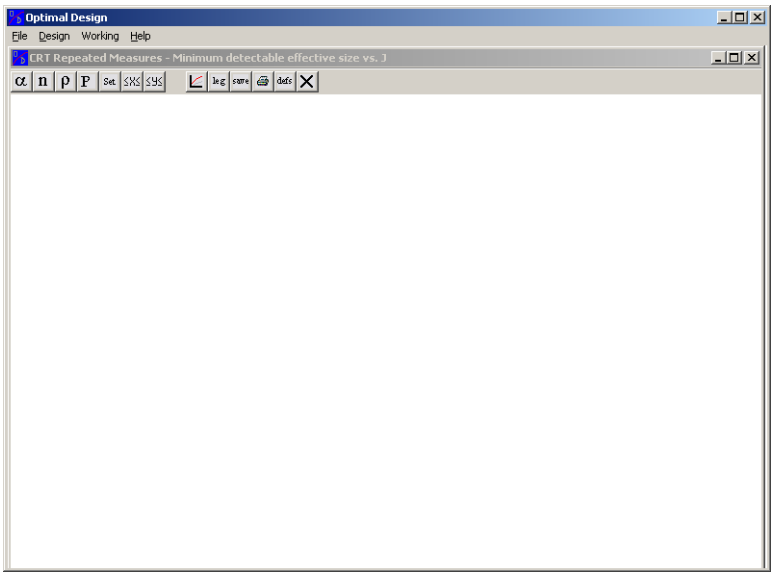


Figure 11.3. Initial screen for cluster randomized trials with repeated measures MDES vs. total number of clusters (*J*).

Step 2: Click on n. Set n(1) = 20

Step 3: Click on ρ. Set ρ (1) = 0.15.

Step 4: Click on *P*. Set *P*(1) = 0.80

Step 5: Click on set. Set f = 1, d = 6, variability of level-1 residual = 1, variability of level-1 coefficient = 1, select polynomial order linear. The final curve appears in Figure 11.4.

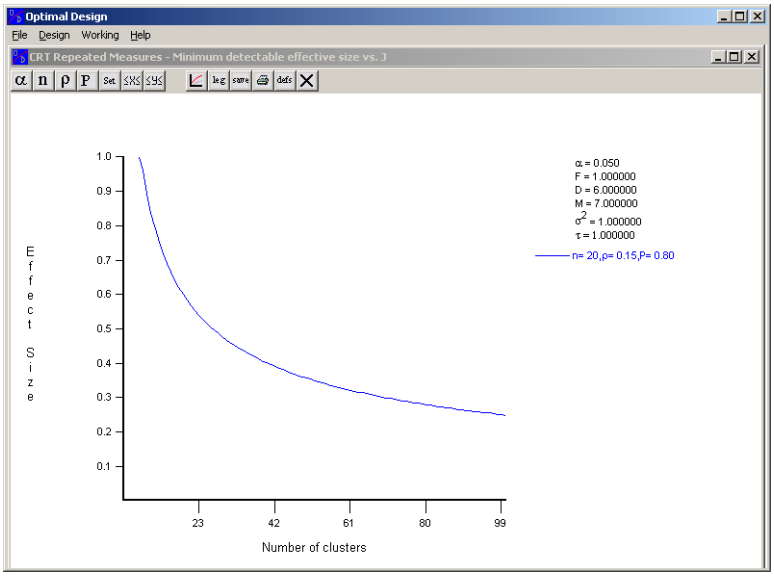


Figure 11.4. Power curve.

Clicking along the trajectory reveals a MDES of approximately 0.32 with *J* = 60.

Section IV: Empirically Based MDES for Cluster Randomized Trials

The Empirically Based MDES is currently available for three specific cluster randomized trials: 2-level CRT, 3-level MSCRT with treatment at Level 2, and 4-level MSCRT with treatment at Level 3. A description of the data and models used to calculate the empirical estimates is available in Appendices A.1 and A.2. Appendix A.1 provides information on the elementary, middle, and high school math and reading data. Appendix A.2 provides information on the Pre-K social-emotional and cognitive data. Because the descriptions of the model and power calculations for each of the three designs were discussed in sections 7, 9, and 10, we refer the reader to these sections for background information. In this section, we provide an overview of the basic screens for the empirically based MDES and a “how to” guide for each of the 3 designs.

12.0 Layout of the Empirically Based MDES

As noted, there are three main options under the Empirically Based MDES:

2-level cluster randomized trial

Multisite cluster randomized trial with treatment at level 2

Multisite cluster randomized trial with treatment at level 3

We first use the main screen for the 2-level CRT to demonstrate the basic setup of the Empirically Based MDES option. The main screen is illustrated by Figure 12.1. The top right corner of the screen has 5 tabs: Level of schooling, Unit of randomization, Outcome/covariate sets, Identify scenarios, and Generate table. To the left of the first four tabs is an indicator of whether or not the user has specified the information indicated. Prior to calculating the MDES and generating the output tables, the user must first specify information for each of the first four tabs. That is, each of the first four tabs must move from Red to Green, or from Not specified to Specified. The main screen for the two types of multisite trials is identical to the 2-level CRT expect that they have one additional tab, effect size variability. This allows the user to treat the sites as fixed or random effects as discussed in sections 14 and 15 for multisite trials.

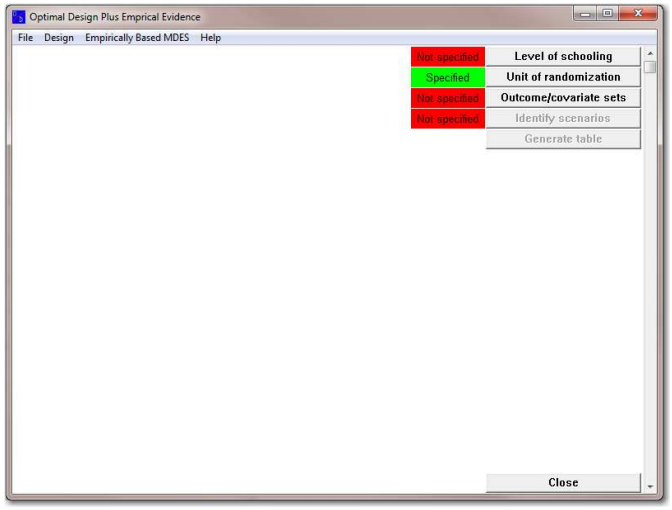


Figure 12.1. Main screen for a 2-level CRT.

Next we describe each of the tabs. There are some differences depending on the particular data available within a design but the main idea behind all the tabs is similar.

Level of schooling. Clicking on this tab opens the popup window in Figure 12.2.

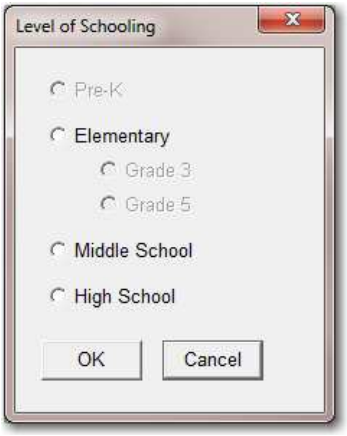


Figure 12.2. Popup window for Level of schooling tab.

This window allows the user to select the level of schooling for the study. If an option is grayed out, then data is not available for a particular grade level or the user must select the overall school level first and then may select the particular grade.

Unit of randomization. Clicking on unit of randomization reveals the popup window in Figure 12.3. This allows the user to specify the unit of randomization. Grayed out options are not available.

Figure 12.3. Popup window for Unit of Randomization.

Outcome/covariate sets. Clicking on this tab opens the popup window in Figure 12.4. The outcome domains available are displayed on the screen. The covariate sets used to calculate the ICCs and R-squares are also listed. Details on the outcomes, covariate sets, and models which generated the values are provided in Appendices A.1 and A.2.

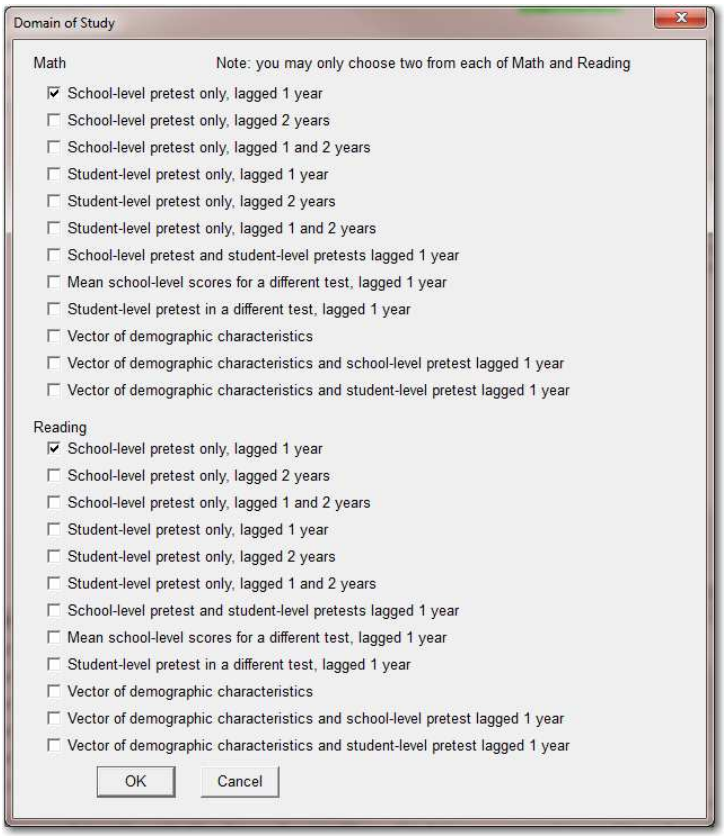


Figure 12.4. Popup window for outcome/covariate sets.

After selecting an outcome and covariate set, the main screen appears with the names of the primary outcomes selected (Figure 12.5). From here, the user selects the box adjacent to the outcome in order to see the plausible values for the ICC and R-squares.

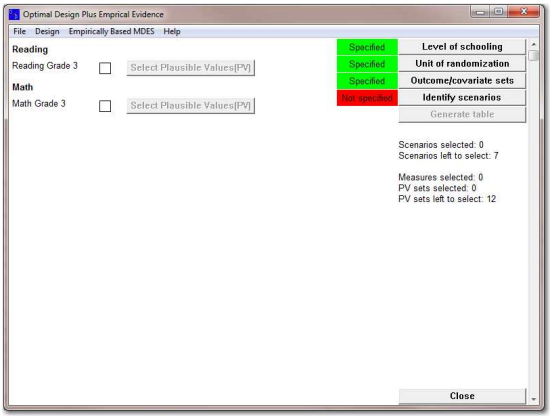


Figure 12.5. Main screen with names of outcomes.

Reading Grade 3, Select Plausible Values (PV). Clicking the box adjacent to Reading Grade 3 reveals the popup window in Figure 12.6. The name of the outcome and covariate set selected appear in the top left. The columns identify what data was used to generate the ICC and

R-squares. The user may select one or more of the columns or may manually enter values for the ICC and R-squares in the final column. For details about the ICCs and R-squares, see Appendix A.1 and A.2.

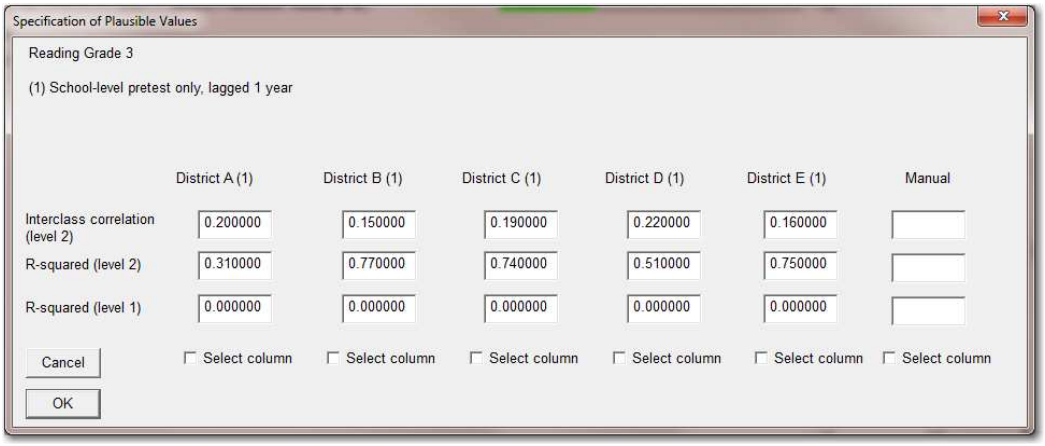


Figure 12.6. Sample popup window for specifying plausible values.

Indentify Scenarios. The identify scenarios button allows the user to specify the sample sizes at all levels. Figure 12.7 shows the popup window for identifying scenarios. The user is allowed to input 7 scenarios at a time. The user is allowed to specify different numbers of treatment and control clusters. It is important to note here that one can specify both integer and non-integer values for the sample size at any level in the analysis. Non-integer values can represent the **harmonic mean** value of the sample size at a given level when this sample size varies within clusters or across sites or blocks. For example, if the number of treatment schools, control schools, or number of students per school varies, one should use their harmonic mean value to specify the corresponding scenario.

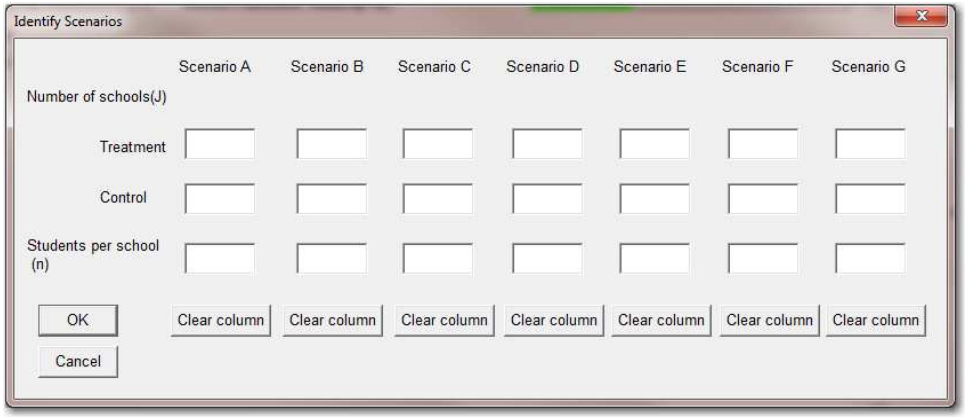


Figure 12.7. Popup window for Identifying Scenarios.

Generate Table. The generate table tab is grayed out until the user has specified all the relevant information. A sample table is provided in Table 12.1. All tables include a title and the power and alpha specifications which define the minimum detectable effect size that is reported. The sample sizes are indicated in the columns while the rows indicate the outcome, the source of the data (i.e. district identified), the covariate set selected, the ICC, and the R-squares. The MDES is in the box. The output table can be printed directly from the screen (File ->Print), saved (File -> Save Copy as), or copy (Ctrl C) and pasted (Ctrl V) into a word document.

Table 12.1. Sample output table for 1-level CRT.

Minimum Detectable Effect Size for Cluster Randomized Trial
Power = 0.80, alpha = 0.05, 2-tail test

	Scenario A J_T=20.0,J_C=20.0,n=50.0
Reading Grade 3 District A (1) ICC_{L2}=0.20 $R^2_{L2,L1}$=0.31,0.00	0.357

(1) School-level pretest only, lagged 1 year

12.1 Empirical Data

Table 12.2 summarizes the level of schooling, unit of randomization, outcome domains, and whether or not there are multiple covariate sets included in the empirical data that is currently available to users in the Empirically Based MDES. Details for the elementary, middle, and high school math and reading data are provided in Appendix A.1. Details for the pre-K: social-emotional and cognitive data are provided in Appendix A.2. We encourage users to read the Appendices to determine the compatibility of the study they are designing with the studies that generated the empirical data used by this program. We expect the program's database to continue to expand and encourage users to check the website frequently for updates.

Table. 12.2 Types of data available in Empirically Based MDES

Design	Level of Schooling	Unit of Randomization	Outcome Domain	Number of Covariate Sets Available
2-level CRT	Elementary, Middle, High School	School	Math, Reading	12
MSCRT with treatment at Level 2	Elementary, Middle, High School	School	Math, Reading	12
MSCRT with treatment at Level 3	Per-K	Pre-K Center	Social-emotional, Cognitive	1

13.0 Two-level Cluster Randomized Trial

In this section, we demonstrate how to use the Empirically Based MDES for planning a 2-level CRT. The underlying models for this design were discussed in section 7. The only difference is that in the empirically based MDES option, we allow covariates at level 1 and 2.

13.1 Example

Suppose a team of researchers develop a new literacy and math program for 3rd graders. The developers of the new program hypothesize that students who participate in the program will have increased reading and math achievement. They plan to test students who participate in the new program (experimental group) and students who participate in the regular program (control group) using a standardized reading test to determine if students using the new program score higher. Suppose that the researchers want to design a cluster randomized trial with students nested within schools where schools are the unit of randomization. They want to access the empirical estimates for calculating the MDES within OD.

The steps for using the Empirically Based MDES follow.

Step 1: Select Empirically Based MDES -> 2-level cluster randomized trial. The main screen is in Figure 13.1.

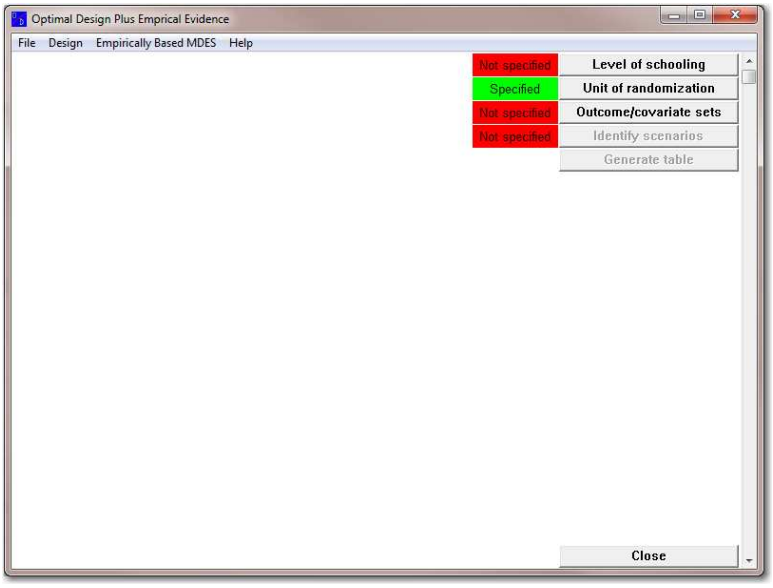


Figure 13.1. Main screen for 2-level CRT.

Step 2: Click on level of schooling and select elementary, grade 3. Figure 13.2 displays the popup window with the selection elementary, grade 3. Click ok.

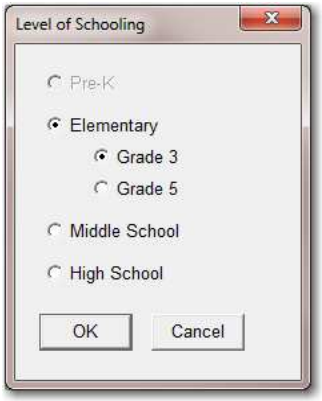


Figure 13.2. Popup window for level of schooling.

Note that unit of randomization is already specified. School is the only option available here so there is no need to click on unit of randomization.

Step 3: Click on outcome/covariate sets. Suppose that the researchers have access to school level pretests lagged one year for reading and math so they want to select these two options. Note that a description of all the covariate sets for reading and math outcomes is provided in Appendix A.1. Figure 13.3 displays the selection. Click ok.

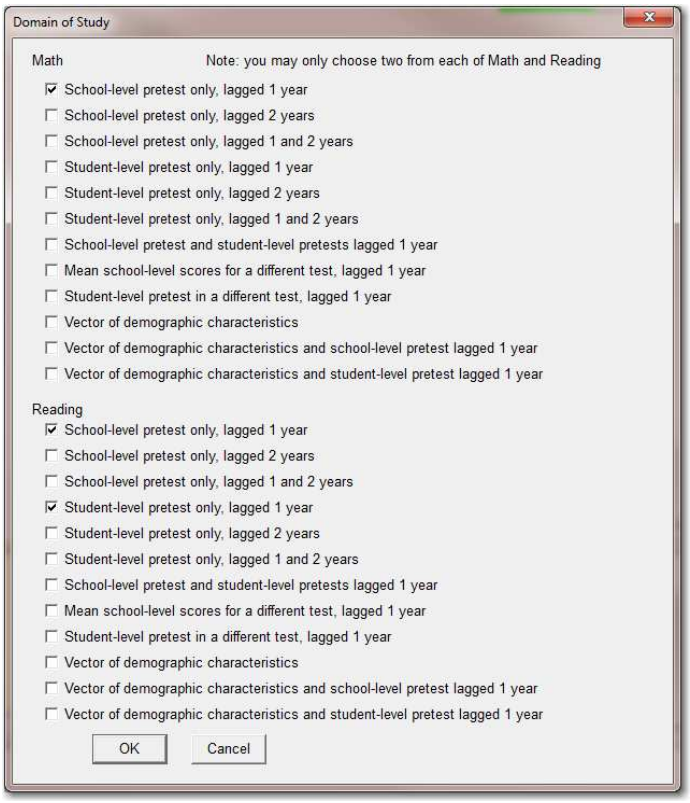


Figure 13.3. Popup window for outcome/covariate sets.

Step 4: Select plausible values for Reading Grade 3: Suppose that District A and B have similar characteristics to the district in the proposed study so select column District A (1) and District B (1) (see appendix A.1 for details on the districts). Selecting two districts will help the user get a sense of the sensitivity of the findings to the varying parameter values. The (1) indicates the covariate set as noted in the top left corner of Figure 13.4. Click Ok after selecting the covariate sets.

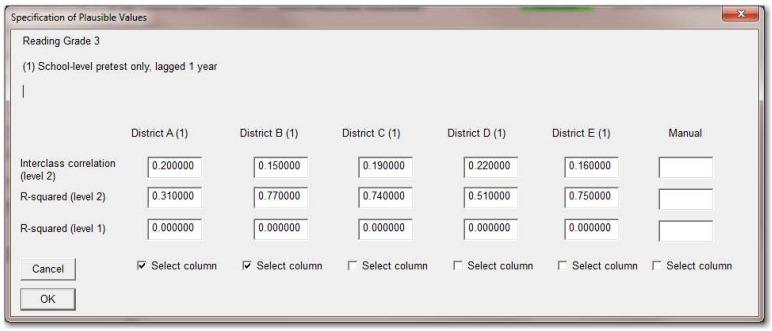


Figure 13.4. Popup window for Specification of Plausible Values.

Step 5: Select Plausible Values for Math Grade 3. The popup window appears for math and again we select the columns for District A and B to build in comparison values. Click ok to continue.

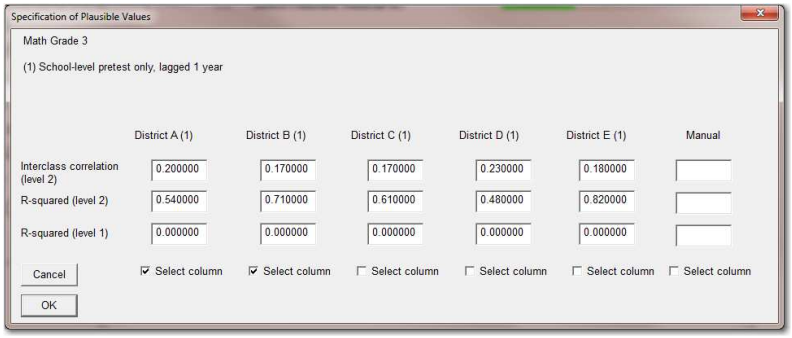


Figure 13.5. Popup window for Specification of Plausible Values.

Step 6: Click on Identify Scenarios. Suppose that the total number of clusters is 40 and the researchers are interested in the MDES assuming equal allocation of clusters and assuming 15 treatment clusters and 25 control clusters. In both cases, assume 50 students per school. Figure 13.6 shows the popup window with the two scenarios specified. Click ok to continue.

Figure 13.6 Popup window for Identify Scenarios.

Step 7: Note that all the tabs are now identified as green or specified. The main screen also shows that 2 scenarios have been selected and 5 more are available. In addition, 2 measures were selected with 4 PV sets specified and 8 remaining. Click on generate table. Table 13.1 displays

the final table. From the table we can see that the MDES is always slightly larger in the imbalanced case (i.e. when the number of treatment schools does not equal the number of control schools). We can also see that the MDES varies depending on the subject and district. Using the parameter values from District A, the MDES for reading and math is 0.357 and 0.299, respectively. For District B, the MDES for reading and math is 0.206 and 0.234, respectively. The MDES from District B was smaller for both domains. In general, the minimum detectable effects sizes are different across districts and outcomes. Thus one must think carefully about the similarities between the district in the proposed study and those used in the calculations as well as the key outcome of interest when it comes time to choosing a final sample design.

Table 13.1 Final output table for 2-level CRT.

Minimum Detectable Effect Size for Cluster Randomized Trial
Power = 0.80, alpha = 0.05, 2-tail test

	Scenario A J_T=20.0,J_C=20.0,n=50.0	Scenario B J_T=15.0,J_C=25.0,n=50.0
Reading Grade 3 District A (1) ICC_{L2}=0.20 $R^2_{L2,L1}$=0.31,0.00	0.357	0.369
Reading Grade 3 District B (2) ICC_{L2}=0.15 $R^2_{L2,L1}$=0.77,0.00	0.206	0.214
Math Grade 3 District A (3) ICC_{L2}=0.20 $R^2_{L2,L1}$=0.54,0.00	0.299	0.309
Math Grade 3 District B (4) ICC_{L2}=0.17 $R^2_{L2,L1}$=0.71,0.00	0.234	0.242

(1) School-level pretest only, lagged 1 year
(2) School-level pretest only, lagged 1 year
(3) School-level pretest only, lagged 1 year
(4) School-level pretest only, lagged 1 year

14.0 Multisite cluster randomized trial with treatment at level 2

In this section, we demonstrate how to use the Empirically Based MDES for planning a MSCRT with treatment at level 2. The underlying models for this design were discussed in section 9. The only difference is that in the empirically based MDES option, we allow covariates at level 1 and 2.

14.1 Example

Let's revisit the example in section 13.1. Suppose a team of researchers develop a new literacy and math program for 3rd graders. The developers of the new program hypothesize that students who participate in the program will have increased reading and math achievement. They plan to test students who participate in the new program (experimental group) and students who participate in the regular program (control group) using a standardized reading test and math test to determine if students using the new program score higher. Suppose that the researchers want to design a multisite cluster randomized trial with students nested within schools which are blocked by district. Schools are the unit of randomization. They want to access the empirical estimates for calculating the MDES within OD.

The steps for using the Empirically Based MDES follow.

Step 1: Select Empirically Based MDES -> multisite cluster randomized trial with treatment at level 2. The main screen is in Figure 14.1. The main screen is similar to the screen presented in section 12 with the addition of the button for effect size variability, which allows the user to specify whether the sites are treated as fixed or random effects.

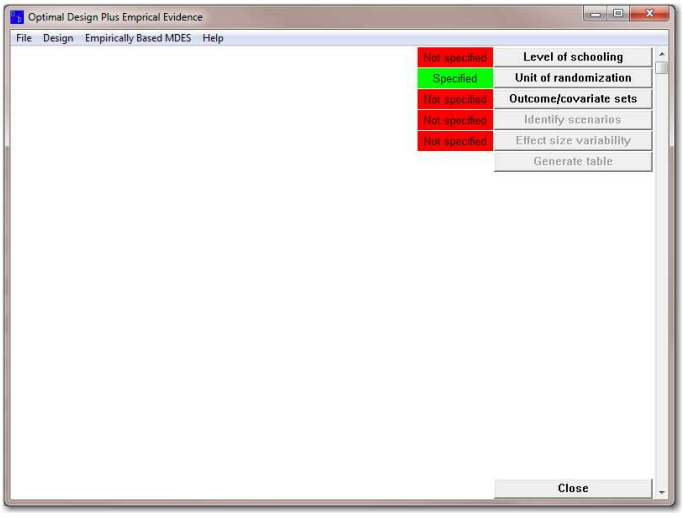


Figure 14.1. Main screen for a MSCRT with treatment at level 2.

Step 2: Click on level of schooling and select elementary, grade 3. Figure 14.2 displays the popup window with the selection elementary, grade 3. Click ok.

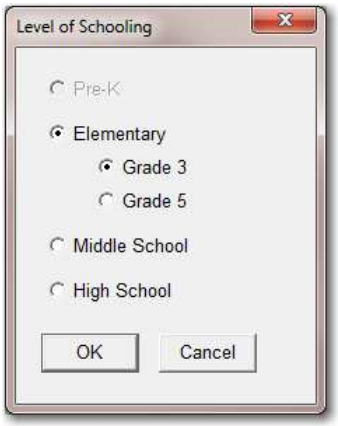


Figure 14.2 Popup window for level of schooling.

Note that unit of randomization is already specified as school so there is no need to click on unit of randomization.

Step 3: Click on outcome/covariate sets. Suppose that the researchers have access to school level pretests lagged one year for reading and math so they want to select this option. Note that a description of all the covariate sets for reading and math is in Appendix A.1. Figure 14.3 displays the selection. Click ok.

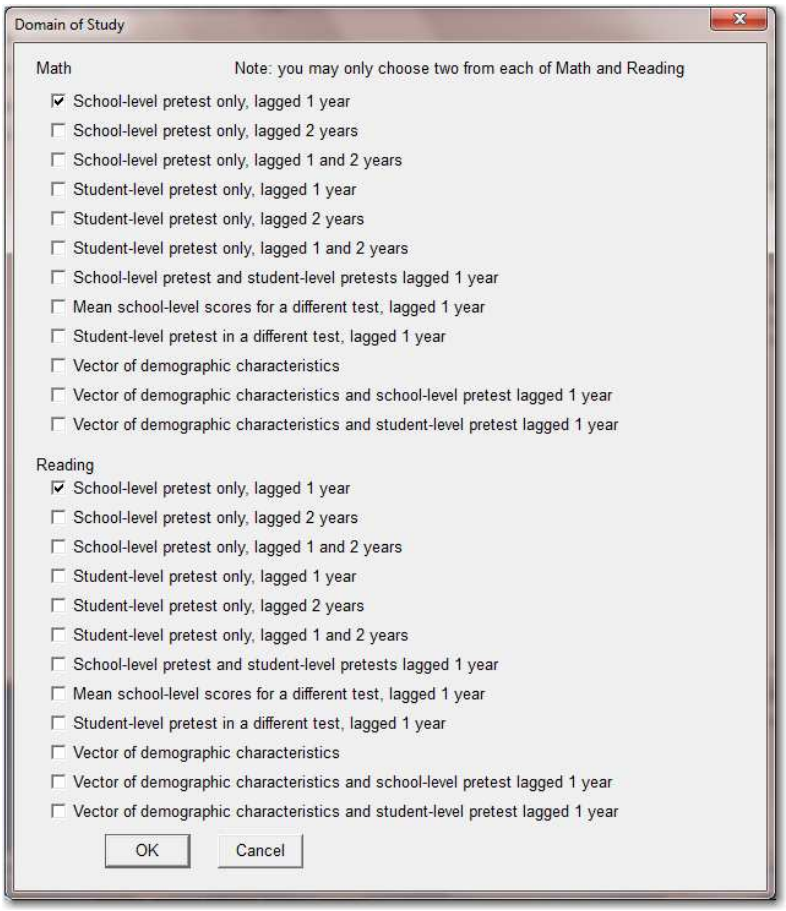


Figure 14.3. Popup window for outcome/covariate sets.

Step 4: Select plausible values for Reading Grade 3. Suppose that District A and B have similar characteristics to the districts in the proposed study so select column District A (1) and District B (1) (see appendix A.1 for details on the districts). Selecting two districts will help the user get a sense of the sensitivity of the findings to the varying parameter values. The (1) indicates the covariate set as noted in the top left corner of the window. Click Ok after selecting the covariate sets.

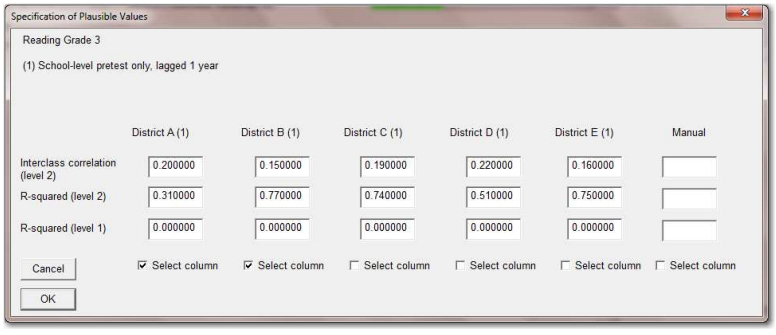


Figure 14.4. Popup window for Specification of Plausible Values.

Step 5: Select plausible values for Math Grade 3. The popup window appears for math and again we select the columns for District A and District B. Click ok to continue.

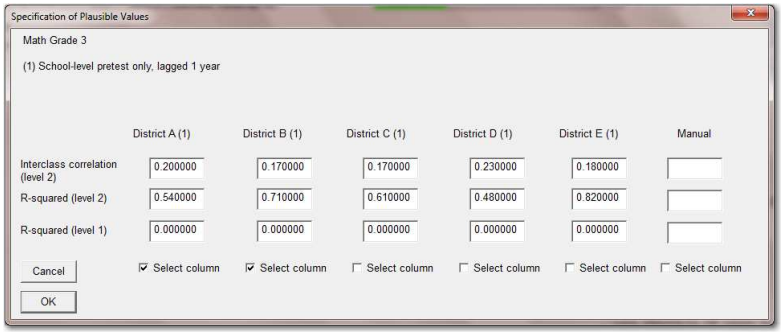


Figure 14.5. Popup window for Specification of Plausible Values.

Step 6: Click on Identify Scenarios. Suppose that the total number of districts or sites is 10. There are 8 schools per district. In scenario A, we specify 4 per treatment and control and assume 50 students per school. In scenario B, we specify 5 per treatment and 3 per control. Figure 13.6 shows the popup window with the scenario specified. Click ok to continue.

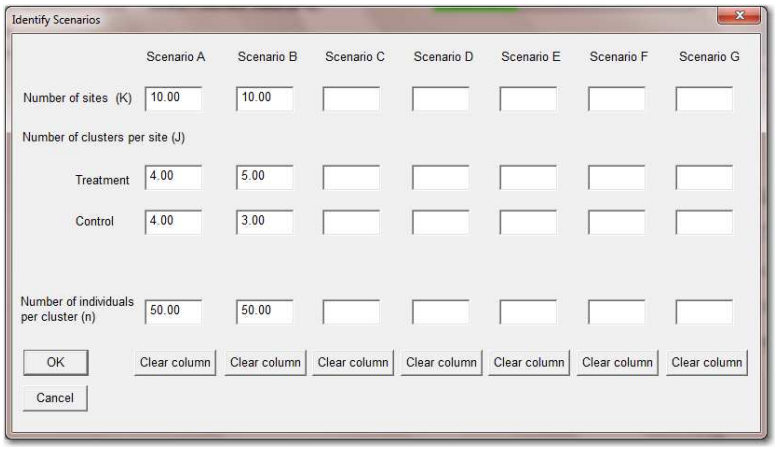
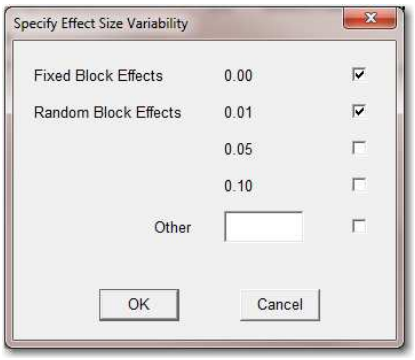


Figure 14.6 Popup window for Identifying Scenarios.

Step 7: Click on effect size variability. The popup window is in Figure 14.7. If the user wants to treat the districts as fixed effects, she selects fixed block effects, 0.00. If the user wants to treat the districts as random effects, she selects one of the options for random effects or enters a manual value. The value corresponds to the effect size variability defined in section 9, or the variability in the treatment effect across districts. In this case, suppose we select fixed block effects and random block effects with 0.01.



14.7. Popup window for effect size variability.

Step 8: Note that all the tabs are now identified as green or specified. The main screen also shows that 1 scenario has been selected and 6 more are available. In addition, 2 measures were selected with 4 PV sets specified and with 8 remaining. Click on generate table. Table 14.1 displays the final table. The table assuming fixed effects is followed by the table assuming random effects. Note that the MDES for fixed effects is always smaller than for random effects. In addition, the MDES is slightly smaller in the balanced case than in the unbalanced case. In general, the minimum detectable effects sizes are quite different across districts and outcomes. Thus one must think carefully about the similarities between the districts in the proposed study and those used in the calculations as well as the key outcome of interest when it comes time to choosing a final sample design.

Table 14.1. Final output table for multisite cluster randomized trial with treatment at level 2.

Minimum Detectable Effect Size for Multisite Cluster Randomized Trial with Treatment at Level 2
Fixed site effects, effect size variability = 0.000
Power = 0.80, alpha = 0.05, 2-tail test

	Scenario A $K=10.0, J_T=4.0, J_C=4.0$ $n=50.0$	Scenario B $K=10.0, J_T=5.0, J_C=3.0$ $n=50.0$
Reading Grade 3 District A (1) $ICC_{L2}=0.20$ $R^2_{L2,L1}=0.31,0.00$	0.251	0.259
Reading Grade 3 District B (2) $ICC_{L2}=0.15$ $R^2_{L2,L1}=0.77,0.00$	0.145	0.150
Math Grade 3 District A (3) $ICC_{L2}=0.20$ $R^2_{L2,L1}=0.54,0.00$	0.210	0.217

Math Grade 3 District B (4) ICC_{L2}=0.17 $R^2_{L2,L1}$=0.71,0.00	0.164	0.170

(1) School-level pretest only, lagged 1 year
(2) School-level pretest only, lagged 1 year
(3) School-level pretest only, lagged 1 year
(4) School-level pretest only, lagged 1 year

Minimum Detectable Effect Size for Multisite Cluster Randomized Trial with Treatment at Level 2
Random site effects, effect size variability = 0.010
Power = 0.80, alpha = 0.05, 2-tail test

	Scenario A K=10.0,J_T=4.0,J_C=4.0 n=50.0	Scenario B K=10.0,J_T=5.0,J_C=3.0 n=50.0
Reading Grade 3 District A (1) ICC_{L2}=0.20 $R^2_{L2,L1}$=0.31,0.00	0.294	0.302
Reading Grade 3 District B (2) ICC_{L2}=0.15 $R^2_{L2,L1}$=0.77,0.00	0.188	0.193
Math Grade 3 District A (3) ICC_{L2}=0.20 $R^2_{L2,L1}$=0.54,0.00	0.252	0.259
Math Grade 3 District B (4) ICC_{L2}=0.17 $R^2_{L2,L1}$=0.71,0.00	0.206	0.212

(1) School-level pretest only, lagged 1 year
(2) School-level pretest only, lagged 1 year
(3) School-level pretest only, lagged 1 year
(4) School-level pretest only, lagged 1 year

15.0 Multisite cluster randomized trial with treatment at level 3

In this section, we demonstrate how to use the Empirically Based MDES for planning a MSCRT with treatment at level 3. The underlying models for this design were discussed in section 10. The only difference is that in the empirically based MDES option, we allow covariates at levels 1, 2, and 3.

15.1 Example

Suppose a team of researchers develop a new preschool curriculum for Head Start students. The developers of the new program hypothesize that students who participate in the program will have improved outcomes in both social-emotional and cognitive domains. They plan to test students who participate in the new program (experimental group) and students who participate in the regular program (control group) using various outcome measures to determine if students using the new program are doing better. Suppose that the researchers want to design a multisite cluster randomized trial with students nested within classes nested within head start centers nested blocked by site. Head start centers are the unit of randomization. They want to access the empirical estimates for calculating the MDES within OD.

The steps for using the Empirically Based MDES follow.

Step 1: Select Empirically Based MDES -> multisite cluster randomized trial with treatment at level 3. The main screen is in Figure 15.1. The main screen is similar to the screen presented in section 12 with the addition of the button for effect size variability, which allows the user to specify whether the sites are treated as fixed or random effects.

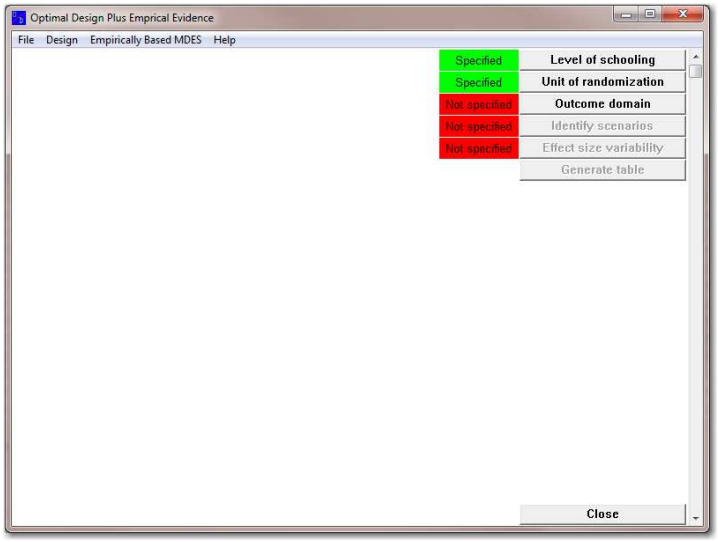


Figure 15.1. Main screen for a MSCRT with treatment at level 3.

Step 2: Click on level of schooling. It is already specified because Pre-K is the only option. Figure 15.2 displays the popup window. Click ok.

Figure 15.2. Popup window for level of schooling.

Step 3: Click on point of randomization. It is already specified because Preschool (Head start center) is the only option available. Figure 15.3 displays the window. Click ok.

Figure 15.3. Popup window for point of randomization.

Step 4: Click on outcome domain and select social-emotional and cognitive (Figure 15.4). Click ok.

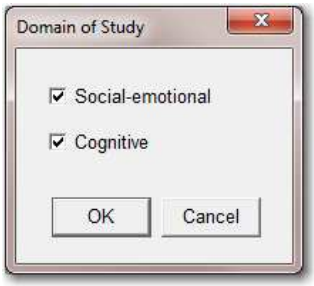


Figure 15.4. Popup window for domain of study.

The screen then appears with the possible measures. A sample of the screen is in Figure 15.5. References for the various measures are included in Appendix A.2.

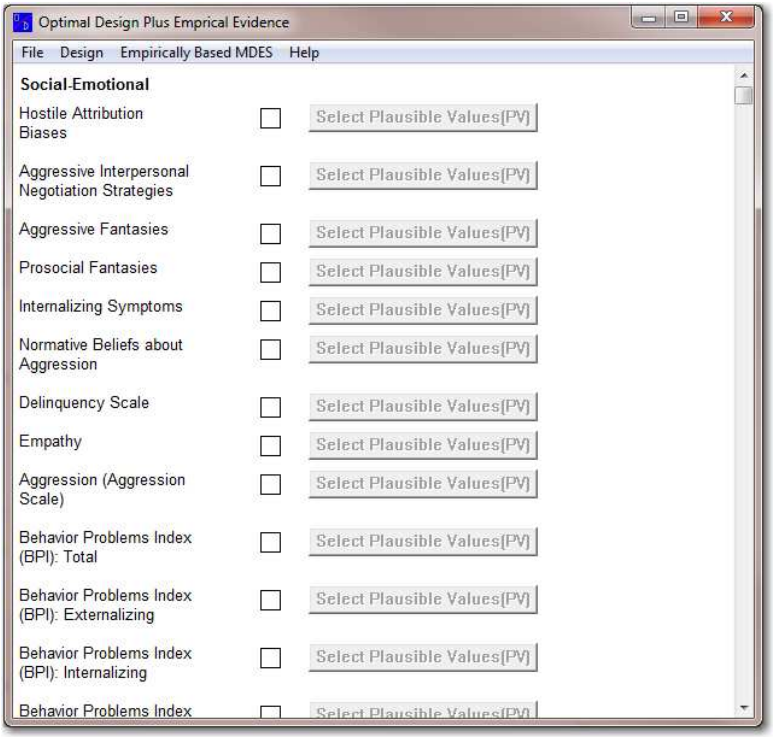


Figure 15.5. Sample screenshot of the social-emotional measures.

Step 5: Suppose we are interested in the measure Peabody Picture Vocabulary Test – III (PPVT). This is the first measure under the domain cognitive so we scroll down to cognitive and select PPVT. The popup window is in Figure 15.6. Each column represents a different study that provides empirical information on the ICCs and R-squares (see Appendix A.2 for details on covariate sets and calculations). Suppose we select the Faces973sSPR99 and the FACES974s PR99 to compare to the MDES from two datasets. Selecting two studies will help the user get a sense of the sensitivity of the findings to the varying parameter values.

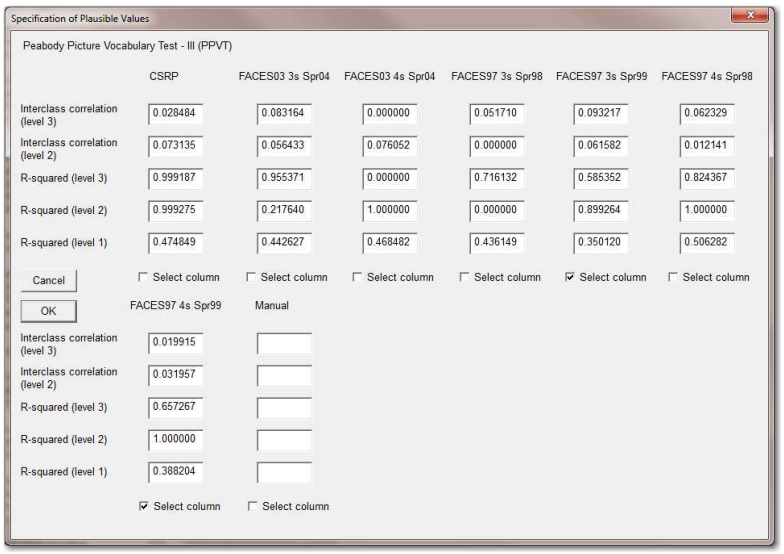


Figure 15.6. Popup window for Specification of Plausible Values.

Step 6: Click on Identify Scenarios. Suppose that the total number of sites is 8 with either 8 or 10 centers per district (evenly split), 4 classrooms per center and 12 children per classroom. The two scenarios are shown in Figure 15.7. Click ok to continue.

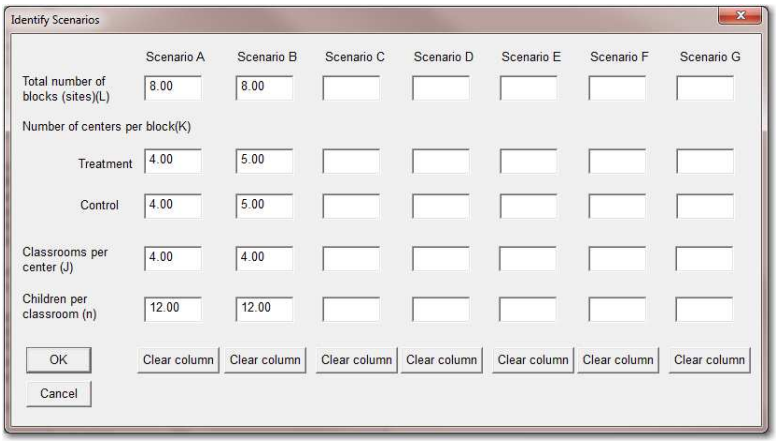
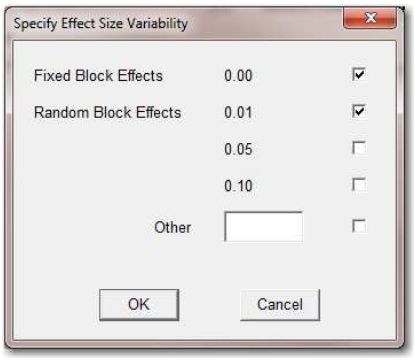


Figure 15.7 Popup window for Identify Scenarios.

Step 7: Click on effect size variability. The popup window is in Figure 15.8. If the user wants to treat the districts as fixed effects, she selects fixed block effects, 0.00. If the user wants to treat the districts as random effects, she selects one of the options for random effects or enters a manual value. The value corresponds to the effect size variability defined in section 10, or the variability in the treatment effect across districts. In this case, suppose we select fixed block effects and random block effects with 0.01.



15.8. Popup window for effect size variability.

Step 8: Note that all the tabs are now identified as green (i.e. specified). The main screen also shows that 2 scenarios have been selected and 5 more are available. In addition, 1 measure was selected with 2 PV sets and there are 10 PV sets remaining. Click on generate table. Table 15.1 displays the final table. The fixed effects table appears first followed by the table assuming random site effects. Note that in this example the minimum detectable effects sizes are quite different for the two measured considered which happens frequently in practice. Thus one must think carefully about one's priorities when it comes time to choosing a final sample design.

Table 15.1 Final output table for multisite cluster randomized trial with treatment at level 3.

Minimum Detectable Effect Size for Multisite Cluster Randomized Trial with treatment at Level 3
Fixed site effects, effect size variability = 0.000
Power = 0.80, alpha = 0.05, 2-tail test

	Scenario A $L=8.0, K_T=4.0, K_C=4.0$ $J=4.0, n=12.0$	Scenario B $L=8.0, K_T=5.0, K_C=3.0$ $J=4.0, n=12.0$
Peabody Picture Vocabulary Test - III (PPVT) FACES97 3s Spr99 $ICC_{L3,L2}=0.09,0.06$ $R^2_{L3,L2,L1}=0.59,0.90,0.35$	0.163	0.169
Peabody Picture Vocabulary Test - III (PPVT) FACES97 4s Spr99 $ICC_{L3,L2}=0.02,0.03$ $R^2_{L3,L2,L1}=0.66,1.00,0.39$	0.099	0.102

Minimum Detectable Effect Size for Multisite Cluster Randomized Trial with treatment at Level 3
Random site effects, effect size variability = 0.010
Power = 0.80, alpha = 0.05, 2-tail test

	Scenario A $L=8.0, K_T=4.0, K_C=4.0$ $J=4.0, n=12.0$	Scenario B $L=8.0, K_T=5.0, K_C=3.0$ $J=4.0, n=12.0$
Peabody Picture Vocabulary Test - III (PPVT) FACES97 3s Spr99 $ICC_{L3,L2}=0.09,0.06$ $R^2_{L3,L2,L1}=0.59,0.90,0.35$	0.219	0.224
Peabody Picture Vocabulary Test - III (PPVT) FACES97 4s Spr99 $ICC_{L3,L2}=0.02,0.03$ $R^2_{L3,L2,L1}=0.66,1.00,0.39$	0.161	0.164

Section V: Optimal Design for cluster randomized trials with binary outcomes

Optimal Design for cluster randomized trials with binary outcomes includes trials where intact groups, or clusters, are randomly assigned to the treatment or control condition. The designs included in this section are the two-level cluster randomized trial (2-level CRT), the three-level cluster randomized trials (3-level CRT), and the multisite cluster randomized trial (MSCRT), a subset of the designs included in Section III, Optimal Design for Cluster Randomized Trials. However, the difference in this section is that the outcome is binary. For example, a study in which the primary outcome is graduation status (yes/no) would require a power analysis for a binary outcome. We describe the conceptual details of each design and provide a “how to” guide for each design in the following 3 chapters.

15.0 Two level cluster randomized trials with a binary outcome

The general design of a 2-level CRT with a binary outcome is the same as a 2-level CRT with a continuous outcome: students nested within schools, or more generally, the level-1 units nested within the level-2 unit. However, the outcome variable is different. For example, the outcome for a study might be whether or not a student drops out of school or whether or not a student drinks alcohol in high school. The variable has only two possibilities so the outcome is binary. Because of the structure of the data, the model for a CRT with a binary outcome is slightly different than the model for a CRT with a continuous outcome. Let's take a closer look at the model.

15.1 The model

The model for a 2-level CRT with a binary outcome can be thought of as an extension of the generalized linear model applied to a multi-level setting. The level-1 model is comprised of three parts: the sampling model, the link function, and the structural model. The level-1 sampling model defines the probability that the event will occur. Let Y_{ij}=1 if an event (often called a "success") occurs and Y_{ij}=0 if not. The sampling model is:

$$Y_{ij} \mid \phi_{ij} \sim B(m_{ij}, \phi_{ij}) \qquad [15.1]$$

for $i \in \{1,2,...,n_j\}$ persons per cluster and for $j \in \{1,2,..., J\}$ clusters;

where m_{ij} is the number of trials for person *i* in cluster *j*; and

ϕ_{ij} is the probability of success for person *i* in cluster *j*.

The expected value and variance of $Y_{ij} \mid \phi_{ij}$ are:

$$E(Y_{ij} \mid \phi_{ij}) = m_{ij}\phi_{ij}$$

$$Var(Y_{ij} \mid \phi_{ij}) = m_{ij}\phi_{ij}(1-\phi_{ij}) \qquad [15.2]$$

Note that in the case of a Bernoulli trial, m_{ij} = 1 so the expected value of $Y_{ij} \mid \phi_{ij}$ reduces to ϕ_{ij} and the variance reduces to $\phi_{ij}(1\text{-}\phi_{ij})$. A common link function for a binary outcome is the logit link:

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right) \qquad [15.3]$$

where η_{ij} is the log odds of success.

The probability of success, the odds of success, and the log odds of success are all related. If the probability of success, ϕ_{ij}, is 0.50, then the odds of success are 0.5/(1-0.5)=1, and the log odds of success is log (1)=0. If the probability of success, ϕ_{ij}, is greater than 0.5, then the odds of success are greater than 1, and the log odds of success is positive. If the probability of success, ϕ_{ij}, is less than 0.5, then the odds of success is less than 1 and the log odds of success is negative.

The third part of the level-1 model is the structural model:

$$\eta_{ij} = \beta_{0j} \qquad [15.4]$$

where β_{0j} is the average log odds of success per cluster *j*.

The level-2 model has the same form as the level-2 model for a 2-level CRT with a continuous outcome. However, the interpretation of the parameters differs because of the logit link function:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \qquad u_{0j} \sim N(0,\tau) \qquad [15.5]$$

where γ_{00} is the average log odds of success across clusters;

γ_{01} is the treatment effect in log odds;

W_j is ½ for treatment and -½ for control;

u_{0j} is the random effect associated with each cluster mean; and

τ is the between cluster variance in log odds.

15.2 Testing the treatment effect

The framework for testing the main effect of treatment in the case of a binary outcome is similar to the case of a continuous outcome. In the model above (equation 15.5), the treatment effect is denoted γ_{01}. It is estimated by:

$$\hat{\gamma}_{01} = \eta_E - \eta_C \qquad [15.6]$$

where η_E is the predicted mean for the experimental group in logs odds and η_C is the predicted mean for the control group in log odds. In a balanced design (equal cluster sizes of size *n*) the variance of $\hat{\gamma}_{01}$ can be approximated by:

$$Var(\hat{\gamma}_{01}) = \frac{4(\tau + \sigma^2/n)}{J} \qquad [15.7]$$

where $\sigma^2 = \left(\frac{1}{\phi_E(1-\phi_E)} + \frac{1}{\phi_C(1-\phi_C)} \right) / 2.$

The test statistic is $\frac{\hat{\gamma}_{01}}{\sqrt{4(\tau + \sigma^2 / n)/J}}$. We use the non-central t-distribution to approximate the power of the test with J-2 degrees of freedom.

15.3 Using the Optimal Design for two-level cluster randomized trials with a binary outcome

The binary outcomes option for the two-level cluster randomized trial is limited to power on the y-axis. The menu for the 2-level CRT with a binary outcome is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 2.

Power on y-axis

Power vs. cluster size (*n*)

Power vs. total number of clusters (*J*)

Power vs. probability of success in treatment group (*phi(E)*)

We present an example below and go through the steps involved in conducting a power analysis for a 2-level CRT with a binary outcome variable.

15.4 Example

Suppose a team of researchers are investigating the effects of a new "Stay in School Campaign." They believe that students that participate in the program are more likely to graduate from high school than students who do not participate in the program. The program targets 12th grade students. The program is implemented at the school level thus we have a nested data structure of students within schools. The outcome for the study is whether or not a student graduates from high school in 4 years. Based on past data, the researchers expect the probability that a student graduates from high school in 4 years to be 0.6 with bounds around this estimate from 0.2 to 0.8. The researchers anticipate the probability that a student graduates to be 0.75 in schools that adopt the new "Stay in School Campaign." They expect to have about 200 students per school. How many schools are required to detect the treatment effect with power = 0.80?

In this example, the total number of clusters, *J*, is unknown. As a result, we want to select

the power vs. number of clusters (*J*) option. This allows the number of clusters to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 2 → Power on y-axis (binary outcome) → Power vs. number of clusters (*J*) as shown in Figure 15.1.

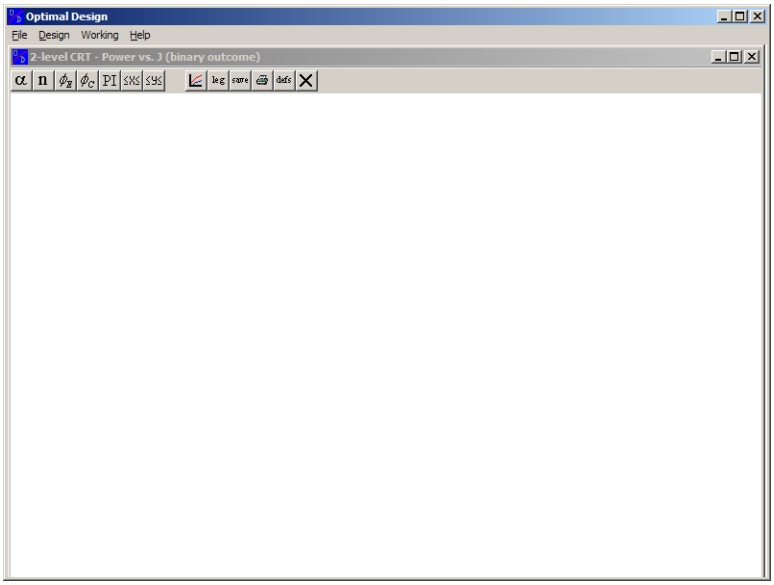


Figure 15.1. Blank screen for 2-level CRT with binary outcomes.

The toolbar at the top includes the parameters required for calculating the power: sample size within cluster (n), the probability of success in the treatment group (ϕ_E), the probability of success in the control group (ϕ_C), and the Plausible Interval for success in the control group. The number of clusters (*J*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 200. The default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on ϕ_E. Set ϕ_E = 0.75. This is the probability of success in the treatment group.

Step 4: Click on PI. Set the lower bound = 0.20 and the upper bound = 0.80. This is the range of plausible values for the probability of success in the control group. Note that ϕ_C must fall within this range. Click on ϕ_C. Set ϕ_C = 0.60. This is the probability of success in the control group.

Step 5: Click on ϕ_C. Set ϕ_C = 0.60. This is the probability of success in the control group.

The resulting power curve is in Figure 12.2.

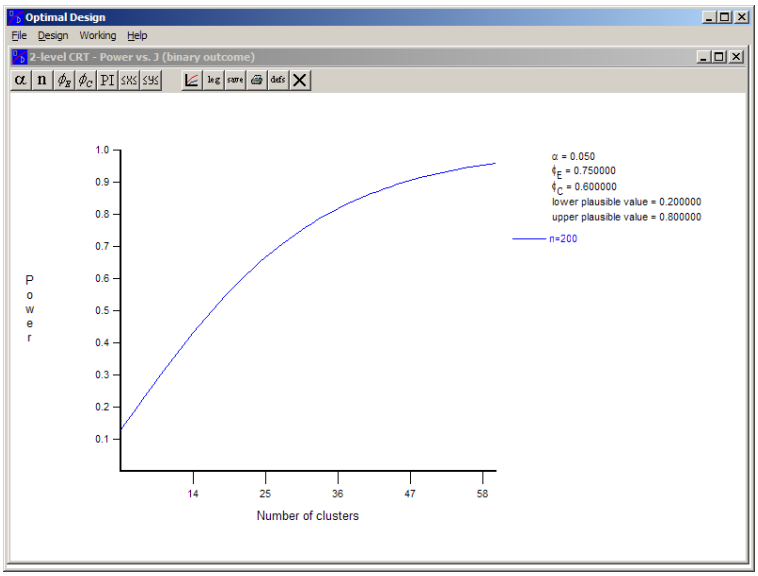


Figure 15.2. Power vs. number of clusters.

Clicking along the power curve, we can see that approximately 36 schools are required for the study.

The example provided in this section placed the total number of clusters on the x-axis. However, the number of persons per cluster or the probability of success in the treatment group could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

16.0 Three level cluster randomized trials with a binary outcome

The general design of a 3-level CRT with a binary outcome is the same as a 3-level CRT with a continuous outcome: for example, students nested within classroom nested within schools, or more generally, the level-1 units nested within the level-2 unit nested within the level-3 unit. However, the outcome variable is binary, that is, there are two possible values the variable can take. For example, the outcome for a study might be whether or not a student drops out of school or whether or not a student drinks alcohol in high school. Because of the structure of the data, the model for a CRT with a binary outcome is more complex than the model for a CRT with a continuous outcome. First we take a closer look at the model.

16.1 The model

The model for a 3-level CRT with a binary outcome can be thought of as an extension of the generalized linear model applied to a multi-level setting. The level-1 model is comprised of three parts: the sampling model, the link function, and the structural model. The level-1 sampling model defines the probability that the event will occur. The sampling model is below:

$$Y_{ijk} \mid \phi_{ijk} \sim B(m_{ijk}, \phi_{ijk}) \qquad [16.1]$$

for $i \in \{1,2,...,n_j\}$ persons per cluster, for $j \in \{1,2,..., J\}$ clusters per site and for $k \in \{1,2,..., K\}$ sites.

where m_{ijk} is the number of trials for person *i* in cluster *j* in site *k*; and

ϕ_{ijk} is the probability of success for person *i* in cluster *j* in site *k*.

The expected value and variance of $Y_{ijk} \mid \phi_{ijk}$ are:

$$E(Y_{ijk} \mid \phi_{ijk}) = m_{ijk}\phi_{ijk}$$

$$Var(Y_{ijk} \mid \phi_{ijk}) = m_{ijk}\phi_{ijk}(1-\phi_{ijk}) \qquad [16.2]$$

Note that in the case of a Bernoulli trial, m_{ijk} = 1 so the expected value of $Y_{ijk} \mid \phi_{ijk}$ reduces to ϕ_{ijk} and the variance reduces to ϕ_{ijk} (1-ϕ_{ijk}). A common link function for a binary outcome is the logit link:

$$\eta_{ijk} = \log\left(\frac{\phi_{ijk}}{1-\phi_{ijk}}\right) \qquad [16.3]$$

where η_{ijk} is the log odds of success.

The probability of success, the odds of success, and the log odds of success are all related. For example, if the probability of success, ϕ_{ijk}, is 0.50, then the odds of success are 0.5/(1-0.5)=1, and the log odds of success is log (1)=0. If the probability of success, ϕ_{ijk}, is greater than 0.5, then the odds of success are greater than 1, and the log odds of success is positive. If the probability of success, ϕ_{ijk}, is less than 0.5, then the odds of success is less than 1 and the log odds of success is negative.

The third part of the level-1 model is the structural model:

$$\eta_{ijk} = \beta_{0\,jk} \qquad [16.4]$$

where $\beta_{0\,jk}$ is the average log odds of success per cluster *j* in site *k*.

The level-2 model takes the same form as the level-2 model for a 3-level CRT with a continuous outcome. However, the interpretation of the parameters differs because of the logit link function:

The level-2 model, or cluster-level model, is:

$$\pi_{0\,jk} = \beta_{00k} + r_{0\,jk} \qquad r_{0\,jk} \sim N(0, \tau_{\pi}) \qquad [16.5]$$

where β_{00k} is the average log odds of success for site k;

$r_{0\,jk}$ is the random effect associated with each cluster; and

τ_{π} is the between-cluster variance in log odds within sites.

The level-3 model, or site-level model, is:

$$\beta_{00k} = \gamma_{000} + \gamma_{001} W_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}}) \qquad [16.6]$$

where γ_{000} is the estimated grand mean in log-odds of success;

γ_{001} is the average treatment effect in log-odds ("main effect of treatment");

W_k is 0.5 for treatment and –0.5 for control;

u_{00k} is the random effect associated with each site mean;

$\tau_{\beta_{00k}}$ is the between-site variance in log-odds.

Note that the randomization in this design occurs at level 3.

16.2 Testing the treatment effect

The framework for testing the main effect of treatment in the case of a binary outcome is very similar to the case of a continuous outcome variable. In the model above (equation 16.6), the treatment effect is denoted γ_{01}. It is estimated by:

$$\hat{\gamma}_{001} = \eta_E - \eta_C \qquad [16.7]$$

where η_E is the predicted mean for the experimental group in logs odds and η_C is the predicted mean for the control group in log odds. The variance of the estimated treatment effect can be approximated by:

$$Var(\hat{\gamma}_{001}) = \frac{4[\tau_{\beta_{00}} + (\tau_\pi + \sigma^2 / n)/J]}{K} \qquad [16.8]$$

where $\sigma^2 = \left(\frac{1}{\phi_E(1-\phi_E)} + \frac{1}{\phi_C(1-\phi_C)} \right) / 2$.

The test statistic is $\frac{\hat{\gamma}_{01}}{\sqrt{(4(\tau_{\beta_{00}} + (\tau_\pi + \sigma^2/n)/J))/K}}$. We use the non-central t-distribution to approximate the power of the test with K-2 degrees of freedom.

16.3 Using the Optimal Design for three level cluster randomized trials with a binary outcome

The menu for the 3-level CRT with a binary outcome is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 3. In this section we focus on binary outcomes as shown below.

Power on y-axis (binary outcomes)

Power vs. cluster size (*n*)

Power vs. number of clusters (*J*)

Power vs. number of sites (*K*)

Power vs. probability of success in treatment group (phi(E))

The options present the power on the y-axis and either the cluster size, number of clusters, or probability of success in the treatment group on the x-axis. We present an example below and

guide the user through the steps for approaching the example via the power determination approach.

16.4 Example

Suppose a team of researchers are investigating the effects of a new "Stay in School Campaign." They believe that students that participate in the program are more likely to graduate from high school than students who do not participate in the program. The program targets 12th grade students. Although the program is adopted school-wide, the program components are delivered in first hour. The researchers suspect there will be differences with respect to the teacher who delivers the program so they are interested in designing a three level study with students nested within teachers nested within schools. The outcome for the study is whether or not a student graduates from high school in 4 years. Based on past data, the researchers expect the probability that a student graduates from high school in 4 years to be 0.6 with an upper and lower bound of 0.2 and 0.8, respectively. The researchers anticipate the probability that a student graduates to be 0.75 in schools that adopt the new "Stay in School Campaign." They expect to have about 25 students per teacher and 6 teachers per school. The researchers also suspect that about two-thirds of the variance is between sites and one-third is between clusters within sites. How many schools are required to detect the treatment effect with power = 0.80?

In the example, the total number of schools, *K*, is unknown. As a result, we want to select the power vs. total number of schools (*K*) option. This allows the number of clusters to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Cluster randomized trials → Treatment at level 3 → Power on y-axis (binary outcome) → Power vs. total number of clusters (*K*) as shown in Figure 16.1.

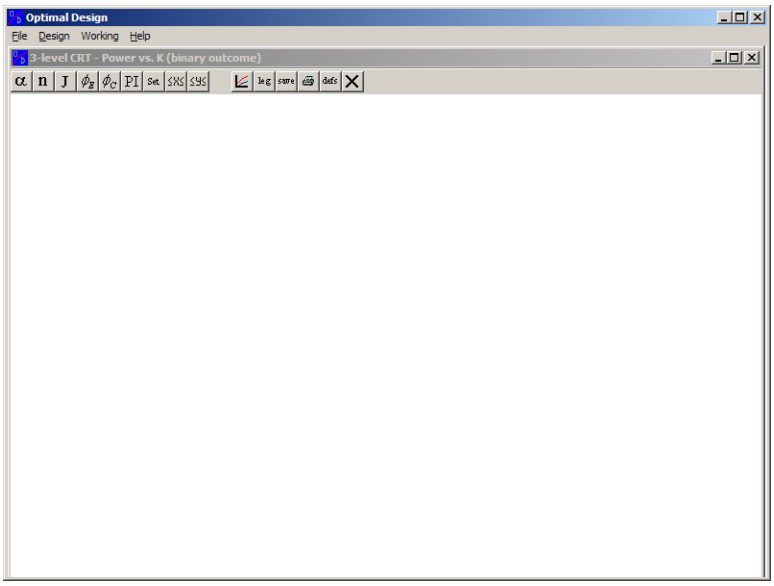


Figure 16.1. Blank screen for 3-level CRT with binary outcomes.

The toolbar at the top includes the parameters required for calculating the power: sample size within cluster (n), the number of clusters per site (*J*), the probability of success in the treatment group (ϕ_E), the probability of success in the control group (ϕ_C), the Plausible Interval for success in the control group, and the set button. The set button asks the user to specify the percent of variability at the cluster level and the site level. The total number of sites (*K*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 25. The default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J*(1) = 6. This the number of clusters per site.

Step 4: Click on ϕ_E. Set ϕ_E = 0.75. This is the probability of success in the treatment group.

Step 4: Click on ϕ_C. Set ϕ_C = 0.60. This is the probability of success in the control group.

Step 5: Click on PI. Set the lower bound = 0.20 and the upper bound = 0.80. This is the range of plausible values for the probability of success in the control group. Note that ϕ_C must fall within this range.

Step 6: Click on set. Specify the percent of variance between clusters within sites. The percent of variance between sites will automatically be calculated as 1 – percent of variance between clusters. Set the percent of variance between clusters = 0.33.

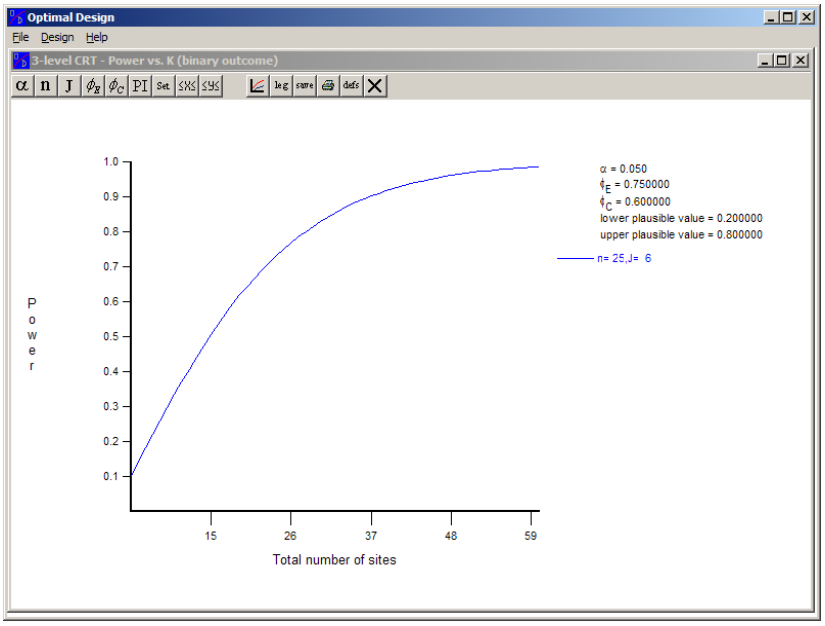


Figure 16.2. Power vs. total number of sites.

Clicking along the power curve, we can see that approximately 28 schools are required for the study, 13 schools in the treatment condition and 13 schools in the control condition.

The example provided in this section placed the total number of sites on the x-axis. However, the number of persons per cluster, the number of clusters per site, or the probability of success in the treatment group could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

17.0 Multisite Cluster Randomized Trials (MSCRT) with Binary Outcomes

The general design of a MSCRT with a binary outcome is the same as a MSCRT with a continuous outcome: for example, students nested within classrooms which are blocked by schools. However, the outcome variable is binary, that is, there are two possible values the variable can take. For example, the outcome for a study might be whether or not a student drops out of school or whether or not a student drinks alcohol in high school. Because of the structure of the data, the model for a MSCRT with a binary outcome is more complex than the model for a CRT with a continuous outcome. Let's take a closer look at the model.

17.1 The model

The model for a MSCRT with a binary outcome and random site effects can be thought of as an extension of the generalized linear model applied to a multi-level setting. The level-1 model is comprised of three parts: the sampling model, the link function, and the structural model. The level-1 sampling model defines the probability that the event will occur. The sampling model is below:

$$Y_{ijk} \mid \phi_{ijk} \sim B(m_{ijk}, \phi_{ijk}) \qquad [17.1]$$

for $i \in \{1,2,...,n_j\}$ persons per cluster, for $j \in \{1,2,..., J\}$ clusters per site and for $k \in \{1,2,..., K\}$ sites.

where m_{ijk} is the number of trials for person *i* in cluster *j* in site *k*; and

ϕ_{ijk} is the probability of success for person *i* in cluster *j* in site *k*.

The expected value and variance of $Y_{ijk} \mid \phi_{ijk}$ are:

$$E(Y_{ijk} \mid \phi_{ijk}) = m_{ijk}\phi_{ijk}$$

$$Var(Y_{ijk} \mid \phi_{ijk}) = m_{ijk}\phi_{ijk}(1-\phi_{ijk}) \qquad [17.2]$$

Note that in the case of a Bernoulli trial, m_{ijk} = 1 so the expected value of $Y_{ijk} \mid \phi_{ijk}$ reduces to ϕ_{ijk} and the variance reduces to ϕ_{ijk} (1-ϕ_{ijk}). A common link function for a binary outcome is the logit link:

$$\eta_{ijk} = \log\left(\frac{\phi_{ijk}}{1-\phi_{ijk}}\right) \qquad [17.3]$$

where η_{ijk} is the log odds of success.

The probability of success, the odds of success, and the log odds of success are all related. For example, if the probability of success, ϕ_{ijk}, is 0.50, then the odds of success are 0.5/(1-0.5)=1, and the log odds of success is log (1)=0. If the probability of success, ϕ_{ijk}, is greater than 0.5, then the odds of success are greater than 1, and the log odds of success is positive. If the probability of success, ϕ_{ijk}, is less than 0.5, then the odds of success is less than 1 and the log odds of success is negative.

The third part of the level-1 model is the structural model:

$$\eta_{ijk} = \beta_{0jk} \qquad [17.4]$$

where β_{0jk} is the average log odds of success per cluster *j* in site *k*.

The level-2 model takes the same form as the level-2 model for a MSCRT with a continuous outcome. However, the interpretation of the parameters differs because of the logit link function:

The level-2 model, or cluster-level model, is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}W_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_{\pi}) \qquad [17.5]$$

where β_{00k} is the average log odds of success for site k;

β_{01k} is the treatment effect at site *k*;

W_{jk} is a treatment contrast indicator, ½ for treatment and -½ for the control;

r_{0jk} is the random effect associated with each cluster; and

τ_{π} is the between-cluster variance in log odds within sites.

The level-3 model, or site-level model, is:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad \text{var}\ (u_{00k}) \sim \tau_{\beta_{00}}$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad \text{var}\ (u_{01k}) \sim \tau_{\beta_{11}} \qquad \text{cov}(u_{00k}, u_{01k}) = \tau_{\beta_{01}} \qquad [17.6]$$

where γ_{000} is the estimated grand mean in log-odds of success;

γ_{010} is the average treatment effect in log-odds;

u_{00k} is the random effect associated with each site mean;

u_{01k} is the random effect associated with each site treatment effect;

$\tau_{\beta_{00}}$ is the between-site variance in log-odds;

$\tau_{\beta_{11}}$ is the between-site variance on the treatment effect in log-odds; and

$\tau_{\beta_{01}}$ is the covariance between site-specific means and site-specific treatment effects.

14.2 Testing the treatment effect

The framework for testing the main effect of treatment in the case of a binary outcome is very similar to the case of a continuous outcome variable. In the model above (Equation 17.6), the treatment effect is denoted γ_{01}. It is estimated by:

$$\hat{\gamma}_{001} = \eta_E - \eta_C \qquad \textbf{[}17.7\text{]}$$

where η_E is the predicted mean for the experimental group in logs odds and η_C is the predicted mean for the control group in log odds. The variance of the estimated treatment effect can be approximated by:

$$Var(\hat{\gamma}_{001}) = \left[\tau_{\beta_{11}} + \left(4(\tau_\pi + \sigma^2 / n)\right)/ J\right]/ K$$

where $\sigma^2 = \left(\frac{1}{\phi_E(1-\phi_E)} + \frac{1}{\phi_C(1-\phi_C)}\right)/2$.

The test statistic is $\frac{\hat{\gamma}_{01}}{\sqrt{(\tau_{\beta_{11}} + [4(\tau_\pi + \sigma^2 / n))/ J]/ K}}$. We use the non-central t-distribution to approximate the power of the test with K-1 degrees of freedom.

17.3 Using the Optimal Design for multisie cluster randomized trials with a binary outcome

The menu for the MSCRT with a binary outcome is shown below and can be found by clicking on the following: Design → Cluster randomized trials with person level outcomes → Multi-site Cluster randomized trials →Treatment at level 2. In this section we focus on binary outcomes as shown below.

Power on y-axis (binary outcomes)

Power vs. cluster size (*n*)

Power vs. number of clusters (*J*)

Power vs. number of sites (*K*)

Power vs. probability of success in treatment group (phi(E))

The options present the power on the y-axis and either the cluster size, number of clusters per site, total number of sites or probability of success in the treatment group on the x-axis. We present an example below and guide the user through the steps for approaching the example.

17.4 Example

Suppose a team of researchers are investigating the effects of a new "Stay in School Campaign." They believe that students that participate in the program are more likely to graduate from high school than students who do not participate in the program. The program targets 12th grade students. The researchers suspect that there are differences between districts so they decide to block on district. That is, within each district, they will randomly assign schools to either the new program or the current program. The outcome for the study is whether or not a student graduates from high school in 4 years. Based on past data, the researchers expect the probability that a student graduates from high school in 4 years to be 0.6 with an upper and lower bound of 0.2 and 0.8, respectively. The researchers anticipate the probability that a student graduates to be 0.75 in schools that adopt the new "Stay in School Campaign." They expect to have about 200 students per school and 6 schools per district. Assuming a small effect size variability, how many districts are required to detect the treatment effect with power = 0.80?

In this example, the total number of districts, *J,* is unknown. As a result, we want to select the power vs. total number of sites (*K*) option. This allows the number of sites to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with person level outcomes → Multisite (or blocked) trials → Treatment at level 2 → Power on y-axis (binary outcome) → Power vs. total number of sites (*K*) as shown in Figure 17.1.

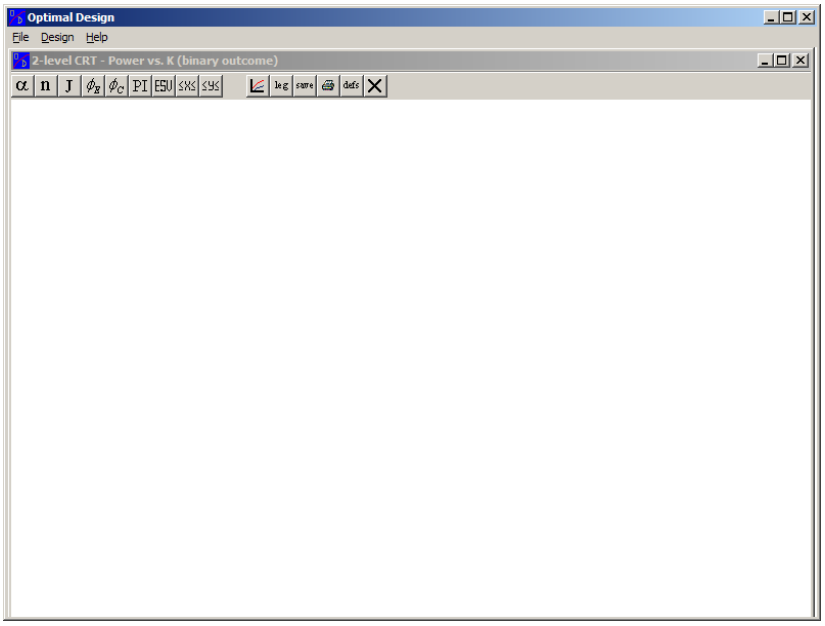


Figure 17.1. Blank screen for power vs. total number of sites.

The toolbar at the top includes the parameters required for calculating the power: sample size within cluster (n), number of clusters per site (*J*), the probability of success in the treatment group (ϕ_E), the probability of success in the control group (ϕ_C),the Plausible Interval for success in the control group and ESV. The ESV specifies the effect size variability, or if the sites are treated as random, how the sites differ with respect to the treatment effect. If the sites are treated as fixed effect, the user should set the ESV to 0. The total number of sites (*K*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on n. Set n(1) = 200. The default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on *J*. Set *J*(1) = 6.

Step 4: Click on ϕ_E. Set ϕ_E = 0.75. This is the probability of success in the treatment group.

Step 5: Click on ϕ_C. Set ϕ_C = 0.60. This is the probability of success in the control group.

Step 6: Click on PI. Set the lower bound = 0.20 and the upper bound = 0.80. This is the range of plausible values for the probability of success in the control group. Note that $\phi_C$must fall within this range.

Step 7: Click on ESV. Set ESV to small. The resulting curve is in Figure 17.2.

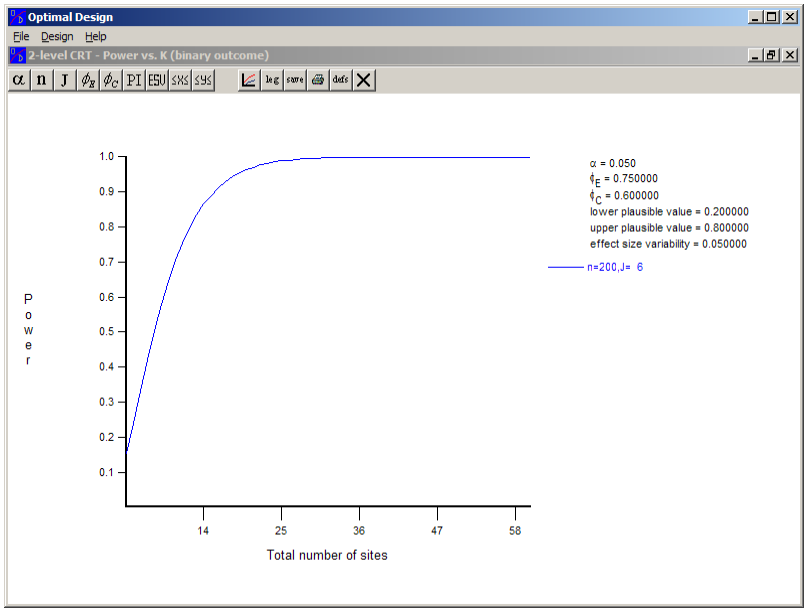


Figure 17.2. Power curve.

Clicking along the power curve, we can see that approximately 13 districts are required for the study with 6 schools per district.

The example provided in this section placed the total number of sites on the x-axis. However, the number of persons per cluster, the number of clusters per site, or the probability of success in the treatment group could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

Section V: Optimal Design for measurement of group processes

Optimal Design for measurement of group processes includes trials where the outcome of interest is measured at the group level, not the individual level. The instruments used to measure the quality of functioning of a group or an organization typically rely on either observational data or interview data. Inherent in both observational data and interview data is measurement error. The magnitude of the measurement errors reduces the reliability of the instrument. Similarly, the reliability is related to the power of a test; the greater the reliability, the greater the power. Thus instruments that exhibit higher levels of reliability are advantageous to researchers planning cluster randomized trials to measure group processes with adequate power. The power analyses in this section account for reliability of measures of groups. The designs included in this section are the two-level cluster randomized trial with a cluster-level outcome, the three-level cluster randomized trial with a cluster-level outcome, and the multisite cluster randomized trial with a cluster-level outcome. We describe the conceptual details of each design and provide a “how to” guide for each design in the following 3 chapters.

18.0 Two-level cluster randomized trials with a cluster-level outcome

Two-level cluster randomized trials with a cluster-level outcome are studies in which the measure is at the group or cluster level and the unit of randomization is also at the group level. For example, suppose a group of researchers are interested in the effect of a classroom intervention on classroom quality. Classrooms are randomly assigned to receive the intervention and observers collect data on classroom quality. The classroom is the unit of measure and the unit of randomization. The unreliability with which the classroom is measured contributes to the power to detect the treatment effect. First, we examine the model in order to determine what affects the power of the study.

18.1 The model

We can think of this as a two level model where level one is a measurement model and level two represents the cluster level. The measurement model captures all sources of measurement error, such as temporal variation, observer variation, individual variation, or item variation. Consider an admittedly over-simplified case in which there one and only one classroom is sampled within each school and where the only source of measurement error is item inconsistency (for now we assume other sources, including rater error and temporal instability, to be null). We also assume each item to be normally distributed about a common classroom mean. Then a level-1 model might be

$$Y_{tj} = \beta_{0j} + e_{tj} \qquad e_{tj} \sim N(0, \sigma^2) \qquad [18.1]$$

for

$t = 1,\ldots,T$ items within a scale

$j = 1,\ldots,J$ clusters

where

β_{0j} is the mean score in cluster j

e_{tj} is the measurement error associated with each item

The level-2 model is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \qquad u_{0j} \sim N(0, \tau) \qquad [18.2]$$

where

γ_{00} is the grand mean

γ_{01} is the treatment effect

u_{0k} is the error associated with each cluster

τ is the between cluster variance.

18.2 Testing the treatment effect

We are interested in the main effect of treatment, γ_{01}, which is estimated by:

$$\hat{\gamma}_{01} = \bar{Y}_E - \bar{Y}_C, \qquad [18.3]$$

where $\bar{Y}_E$ is the mean for the experimental group and $\bar{Y}_C$ is the mean for the control group. When each treatment has an equal number, *J*/2, of clusters, the variance of the main effect of treatment is:

$$Var(\hat{\gamma}_{01}) = \frac{4(\tau + \sigma^2 / T)}{J} \qquad [18.4]$$

where *T* is the number of items within the scale and *J* is the total number of clusters.

We are interested in testing the null hypothesis of no treatment effect.

$H_0 : \gamma_{01} = 0$ vs. $H_1 : \gamma_{01} \neq 0$,

The test statistic is an *F* statistic. Assuming there is a difference in groups, the *F* test follows a non-central *F* distribution, $F(1, J\text{-}2; \lambda)$. Below is the noncentrality parameter for the test, λ, which is the ratio of the squared-treatment effect to the variance of the treatment effect estimate.

$$\lambda = \frac{\gamma_{01}^2}{\hat{Var}(\gamma_{01})} = \frac{K\gamma_{001}^2}{4(\tau + \sigma^2 / T)/J} \qquad [18.5]$$

18.3Standardized Notation

For studies measuring group processes, we standardize the model differently than for studies measuring individuals. The first difference is in the effect size. We define the standardized effect size below:

$$\delta = \frac{\gamma_{01}}{\sqrt{\tau}} \qquad [18.6]$$

Note that the effect size is divided by the square root of the level-2 variance. This differs from the 2-level CRT notation for individual-level outcomes, which divides the effect size by the

square root of the sum of the level-one and level-two variance. The reason this effect size is divided by only the level-2 variance is because we are measuring a level-2 process. We are looking at change in group processes so we only want to use the between group variance and not the measurement error to standardize the value. In other words, standardization occurs at the level of randomization.

The second main difference is that we use reliability instead of the intra-class correlation for purposes of assessing power. We assume the unit of randomization is the same as the unit of measurement and define the reliability at the cluster level in Equation 18.7.

$$reliab_{L2} = \frac{\tau}{\tau + \sigma^2 / T} \qquad [18.7]$$

The reliability is similar to the intra-class correlation except that it adjusts for the number of items within a scale.

The non-centrality parameter can also be defined in terms of the effect size in equation 18.6 and the level-2 reliability in equation 18.7. The non-centrality parameter is:

$$\lambda = \frac{(reliab_{L2})\delta^2}{4 / J} \qquad [18.8]$$

Looking at the noncentrality parameter, we can see that increasing the reliability increases the power. Because the reliability is a function of the variance and the number of items in a scale and the variance is typically not under the control of the researcher, increasing the number of items in a scale is one method for increasing the reliability and hence the power of the study. In addition to the reliability, increasing the number of clusters also increases the power. Finally, larger effect sizes result in greater power. However, the size of the effect is often determined by the phenomenon under investigation, not by the researcher.

We can generalize the example to any case in which group quality is measured with reliability denoted as *reliab*, including cases in which the reliab takes into account multiple sources of error, including temporal instability and rater inconsistency, for example.

18.4 Using the Optimal Design for two-level cluster randomized trials with a cluster-level covariate

This section focuses on how to use the Optimal Design software to design a two-level cluster randomized trial with a cluster-level outcome. The menu for the 2-level CRT with a cluster level covariate is shown below and can be found by clicking on the following: Design →

Cluster randomized trials with cluster-level outcomes (measurement of group processes) → Cluster randomized trials → Treatment at level 2.

Power on y-axis

- Power vs. number of clusters (*J*)
- Power vs. effect size (delta)
- Power vs. cluster-level reliability (Rel(L2))

MDES on y-axis

- MDES vs. number of clusters (*J*)
- MDES vs. power (*P*)
- MDES vs. cluster-level reliability (Rel(L2))

The first set of options present the power on the y-axis and either the number of clusters, effect size, or the cluster-level reliability on the x-axis. The second set of options present the MDES on the y-axis and either the number of clusters, power, or cluster-level reliability on the x-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

18.5 Example

Suppose a team of researchers want to measure the impact of an intervention on the quality of functioning at pre-school sites. The founders of the intervention propose that the overall quality of functioning will increase with participation in their program. They plan assess the quality of the pre-schools using an observational instrument. The researchers plan to randomly assign pre-school sites to either the treatment or control, hence they have a cluster randomized trial. Section 18.6 presents a scenario in which the power determination approach for conducting a power analysis is most applicable and the details of how to do the power analysis using OD. Section 18.7 presents a scenario in which the effect size approach is most applicable and the details of how to do the power analysis using OD.

18.5 Power determination approach for conducting a power analysis

Based on previous studies that use the same observational instrument, the researchers estimate the pre-school level reliability equals 0.75 and want to be able to detect a minimum effect size of 1.0. How many pre-schools are required in order to achieve power equal to 0.80?

In Scenario 1, the number of clusters, *J*, is unknown. As a result, we want to select the power vs. number of clusters (*J*) option. This allows the number of clusters to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with cluster level outcomes (measurement of group processes) → Cluster randomized trials → Treatment at level 2 → Power on y-axis → Power vs. number of clusters (*J*) as shown in Figure 18.1.

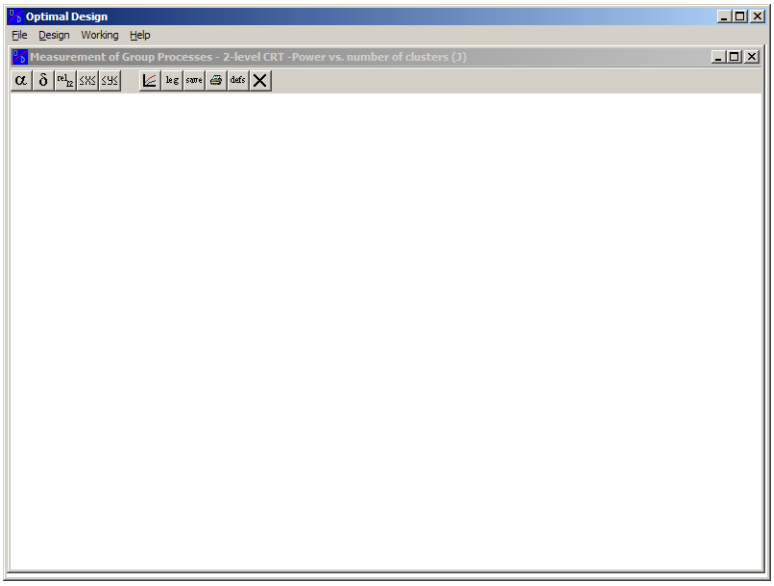


Figure 18.1. Initial blank screen for power vs. number of clusters (*J*).

The toolbar at the top includes the parameters required for calculating the power: the effect size (δ) and the reliability of the cluster-level covariate (relL2). The number of clusters (*J*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on δ. Set delta(1) = 1.0. Recall that delta in the case of a 2-level CRT for a cluster level covariate is defined as $\delta = \frac{\gamma}{\sqrt{\tau}}$ and delta in a 2-level CRT with an individual level covariate is defined as $\delta = \frac{\gamma}{\sqrt{\tau + \sigma^2}}$. The default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 3: Click on relL2. Set relL2 (1) = 0.75. The resulting power curve is in Figure 18.2.

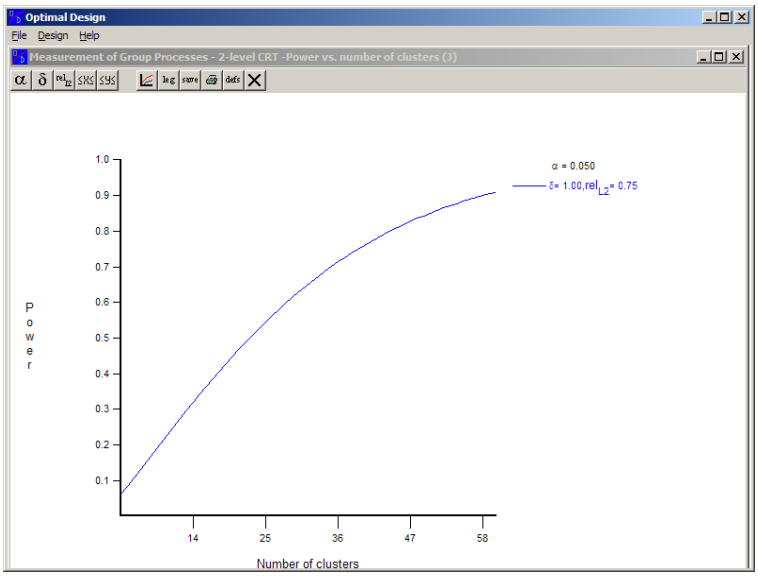


Figure 18.2. Power curve.

Clicking along the power curve, we can see that approximately 44 total clusters are required for the study, 22 in the treatment condition and 22 in the control condition.

The example provided in this section placed the sample size on the x-axis. However, the reliability or the effect size could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

18.6 Effect size approach for conducting a power analysis

Based on previous studies that use the same observational instrument, the researchers estimate the pre-school level reliability equals 0.75. They have secured 40 classrooms, 20 in the treatment and 20 in the control. What is the MDES with power = 0.80?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. number of clusters (*J*). This will allow the user to see how the MDES changes as a function of the total number of clusters holding the power constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with cluster level outcomes (measurement for group processes) → Cluster randomized trials → Treatment at level 2 → MDES on y-axis → MDES vs. number of clusters (*J*) as shown in Figure 18.3.

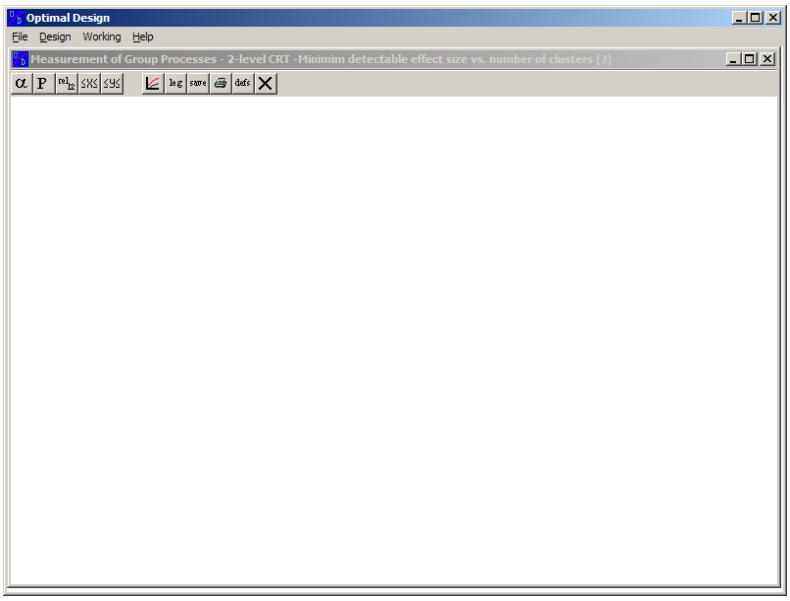


Figure 18.3. Blank screen for MDES vs. number of clusters.

Step 2: Click on *P*. Set *P*(1) = .80.

Step 3: Click on relL2. Set Reliab(l2)(2) = 0.75. The resulting power curve appears in Figure 18.4.

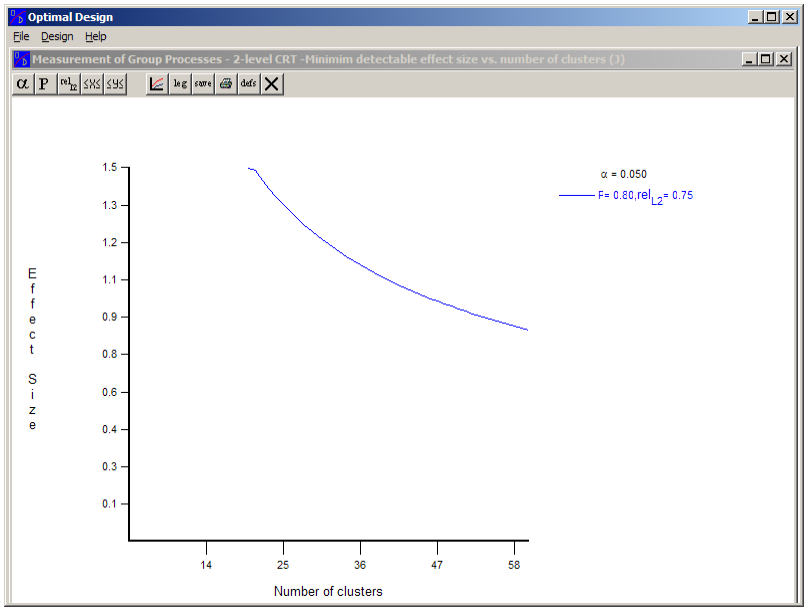


Figure 18.4. Power vs. number of clusters.

Clicking on the trajectory reveals that with 40 clusters, the MDES is approximately 1.06.

The examples in this section are meant to provide at guide to users for how to use the 2-level CRT with a cluster level outcome. We described Power vs. number of clusters (*J*) and MDES vs. number of clusters (*J*). The other options function similarly, and simply place a

different parameter on the x-axis. The choice of which module is most appropriate depends on the unknown parameters. However, all modules yield the same results if identical parameters are used so the choice depends on what module is most closely aligned with the known and unknown parameters in a study.

19.0 Three-level cluster randomized trials with a cluster-level outcome

A three level trial with randomization at level three is a commonly used design. For example, imagine an evaluation for a new elementary math program. Schools are randomly assigned to either the new program or their regular program. Within each school, all the classrooms adopt the new program. The effects of adopting the new program are evaluated by the quality of functioning of a classroom. This means that the outcome is being measured at the level two/cluster level. The measurement usually relies on either observational or interview data, which means it includes measurement error. The measurement error as well as the nested structure of the data has implications for statistical power. First, we take a closer look at the models.

19.1 The model

We can think of this as a three level model where level one is the measurement model. the level-1 model is a measurement model. We again begin with a simplified case in which item inconsistency is the only source of measurement error:

$$Y_{tjk} = \pi_{0jk} + e_{tjk} \qquad e_{tjk} \sim N(0, \sigma^2) \qquad [19.1]$$

for

$t = 1,..., T$ items within a scale

$j = 1,..., J$ clusters (classroom in this case)

$k = 1,..., K$ sites (school in this case)

where

Y_{tjk} is the observed outcome for item t in cluster j in site k

π_{0jk} is the mean for cluster j in site k

e_{tjk} is random error associated with the item

σ^2 is the measurement variance.

The level-2 model is

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi) \qquad [19.2]$$

where

β_{00k} is the mean for site *k*

r_{0jk} is random error associated with the cluster

τ_{π} is the variance between clusters within sites.

The level-3 model is

$$\beta_{00k} = \gamma_{000} + \gamma_{001}W_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00k}}) \qquad [19.3]$$

where

γ_{000} is the grand mean

γ_{001} is the treatment effect ("main effect of treatment")

W_k is a treatment contrast indicator, 0.5 for treatment and –0.5 for control

u_{00k} is the random error associated with each site mean

$\tau_{\beta_{00k}}$ is the residual variance between site means.

19.2 Testing the treatment effect

In the model above, the treatment effect is estimated at level-3 and is denoted γ_{001}. Given a balanced design, it is estimated by:

$$\hat{\gamma}_{001} = \bar{Y}_E - \bar{Y}_C \qquad [19.4]$$

where $\bar{Y}_E$ is the mean for the experimental group

$\bar{Y}_C$ is the mean for the control group.

Because of the nested structure of the data, we sum over clusters and sites in order to estimate the treatment effect. The variance of the estimated treatment effect combines the variance at all three levels, the variance between-site means, $\tau_{\beta_{00}}$, the within-site or between-cluster variance, τ_{π}, and the within-cluster measurement variance, σ^2.

Assuming balanced allocation of clusters to treatment and control, the variance of the treatment effect is estimated by:

$$Var(\hat{\gamma}_{001}) = \frac{4[\tau_{\beta_{00}} + (\tau_{\pi} + \sigma^2/T)/J]}{K} \qquad [19.5]$$

where T is the number of items within the scale, J is number of cluster per site and K is the total number of sites.

To test the null hypothesis of no treatment effect, we use an F statistic.

$H_0 : \gamma_{001} = 0$ vs. $H_1 : \gamma_{001} \neq 0$,

The F test follows a non-central F distribution, $F(1, K\text{-}2; \lambda)$. Recall that the noncentrality parameter, λ, is a ratio of the squared-treatment effect to the variance of the treatment effect estimate. Below is the noncentrality parameter for the test.

$$\lambda = \frac{\gamma_{001}^2}{\hat{Var}(\gamma_{001})} = \frac{K\gamma_{001}^2}{4[\tau_{\beta_{00}} + (\tau_\pi + \sigma^2 / T)/J]} \qquad [19.6]$$

19.3 Standardized notation

To reinforce the concept of the measurement model, we standardize the model differently:

Here, we define the standardized effect size below:

$$\delta = \frac{\gamma_{001}}{\sqrt{\tau_\pi + \tau_{\beta_{00}}}} \qquad [19.7]$$

Note that the effect size is divided by the square root of the level-2 and level-3 variance. This differs from the three-level model when measurement occurs at the individual level, which divides the effect size by the square root of the sum of the level-1, level-2 and level-3 variance. The reason this effect size is divided by only the level-2 and level-3 variance is because we are measuring a level-2 process. So we do not use measurement error to standardize the treatment effect.

Since we are measuring the outcome at the classroom level, the measurement reliability is the key factor which would affect the power of the study. The measurement reliability at the classroom is:

$$reliab_{L2} = \frac{\tau_\pi}{\tau_\pi + \sigma^2 / T} \qquad [19.8]$$

Also, the percentage of variance of average treatment effect between schools is:

$$\rho_{level3} = \frac{\tau_{\beta_{00}}}{\tau_\pi + \tau_{\beta_{00}}} \qquad [19.9]$$

Notice that here the percentage of variance between schools to the total variance is different from the intra-class correlation ρ_{level3} in the 3-level CRT because the first level is a measurement model whose variance is measurement errors.

In standardized notation, the non-centrality parameter, λ, can be rewritten as:

$$\lambda = \frac{K\delta^2}{4\left(\rho_{level3} + \frac{1-\rho_{level3}}{J * reliab_{L2}}\right)} \qquad [19.10]$$

Recall that increasing the noncentrality parameter increases the power to detect the treatment effect. The size of the treatment effect is often based on theory, past studies, or a pilot study which means the researcher cannot inflate the size of the treatment effect to increase power without decreasing the theoretical or practical conclusions of the study. Equation 10 reveals that decreasing the variance between sites will increase the power. The OD allows the user to enter a site-level covariate which can reduce $\tau_{\beta_{00}}$. However, the τ_{π}, and σ^2 are not under the control of the researcher. Equation 16.10 also reveals that increasing the measurement reliability increases the power. For example, we can either increase the number of items during measuring, T to increase the level two reliability. Indeed, from equation 16.10, we can see that increasing the number of sites *K*, or number of clusters, *J*, can also increase the power.

19.4 Using the Optimal Design for three-level cluster randomized trials with a cluster-level covariate

This section focuses on how to use the Optimal Design software to design a three-level cluster randomized trial with a cluster-level outcome. The menu for the 3-level CRT with a cluster level outcome is shown below and can be found by clicking on the following: Design -> Cluster randomized trials with cluster-level outcomes (measurement of group processes) -> Cluster randomized trials -> Treatment at level 3.

Power on y-axis

Power vs. number of level-2 units (*J*)

Power vs. number of level-3 units (*K*)

Power vs. effect size (delta)

Power vs. intraclass correlation at level-3 (rho)

Power vs. reliability at level-2 (Rel(L2))

MDES on y-axis

Power vs. number of level-2 units (*J*)

Power vs. number of level-3 units (*K*)

Power vs. power (*P*)

Power vs. intraclass correlation at level-3 (rho)

Power vs. reliability at level-2 (Rel(L2))

The first set of options present the power on the y-axis and the second set of options present the MDES on the y-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

19.5 Example

Suppose a team of researchers is interested in the effect of a new comprehensive school reform (CSR) on the organization of a classroom. The CSR is implemented at the school level. Schools are randomly assigned to either the CSR or their regular teaching methods. The organization of the school will be measured via an observational instrument. Data from other studies were obtained prior to this study. This enabled the researchers to estimate the reliability at the school level. Their estimate took into account temporal variability and observer variability since schools were assessed at different times and by different people. Assume the reliability at the school level was 0.75, and intraclass correlation is 0.15. Section 19.6 presents a scenario in which the power determination approach for conducting a power analysis is most applicable and the details of how to do the power analysis using OD. Section 19.7 presents a scenario in which the effect size approach is most applicable and the details of how to do the power analysis using OD.

19.6 Power determination approach for conducting a power analysis

The researchers secure 14 classrooms per school. They are interested in an effect size of 0.75. How many schools are required in order to achieve power equal to 0.80?

In Scenario 1, the number of level-3 units, *K*, is unknown. As a result, we want to select the power vs. number of level-3 units (*K*) option. This allows the number of schools to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with cluster level outcomes (measurement of group processes) → Cluster randomized trials → Treatment at level 3 → Power on y-axis →Power vs. number of level-3 units(*K)* as shown in Figure 19.1.

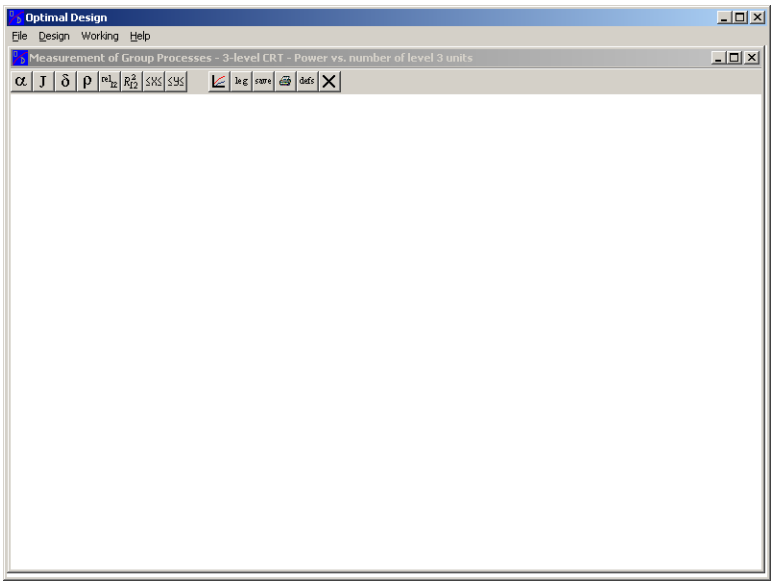


Figure 19.1. Initial blank screen of Power vs. level-3 units (*K*).

Step 2: Click on *J*. Set *J*(1) = 8.

Step 3: Click on δ. Set delta(1) = 0.75.

Step 4: Click on ρ. Set rho (1) = 0.15.

Step 5: Click on relL2. Set relL2 (1) = 0.75. The power curve is in Figure 19.2.

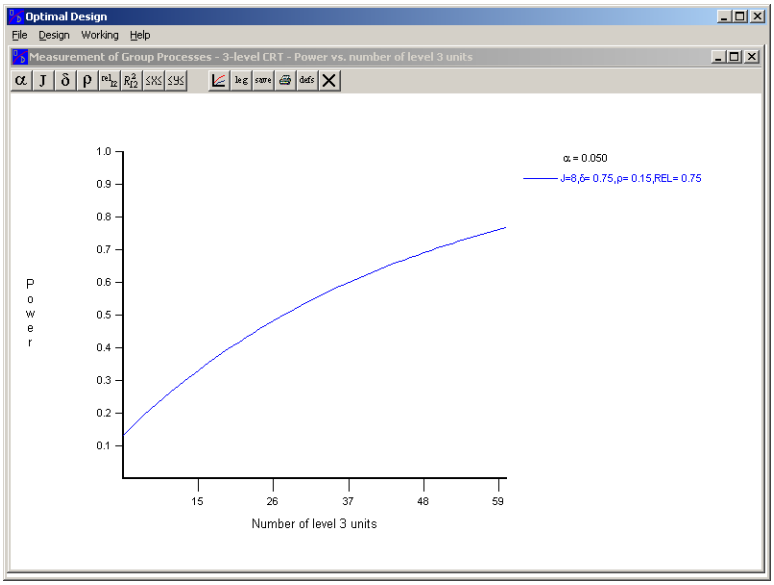


Figure 16.2. Power curve.

Clicking along the power curve, we can see that approximately 54 total clusters are required for the study, 27 in the treatment condition and 27 in the control condition.

The example provided in this section placed the sample size on the x-axis. However, the other parameters could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

19.6 Effect size approach for conducting a power analysis

The researchers have secured 40 schools, 20 in the treatment and 20 in the control and 10 classes per school. What is the MDES with power = 0.80?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. number of level-3 units (*K*). This will allow the user to see how the MDES changes as a function of the total number of schools holding the power constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with cluster level outcomes (measurement for group processes) → Cluster randomized trials → Treatment at level 3 → MDES on y-axis → MDES vs. number of level-3 units (*K*) as shown in Figure 19.3.

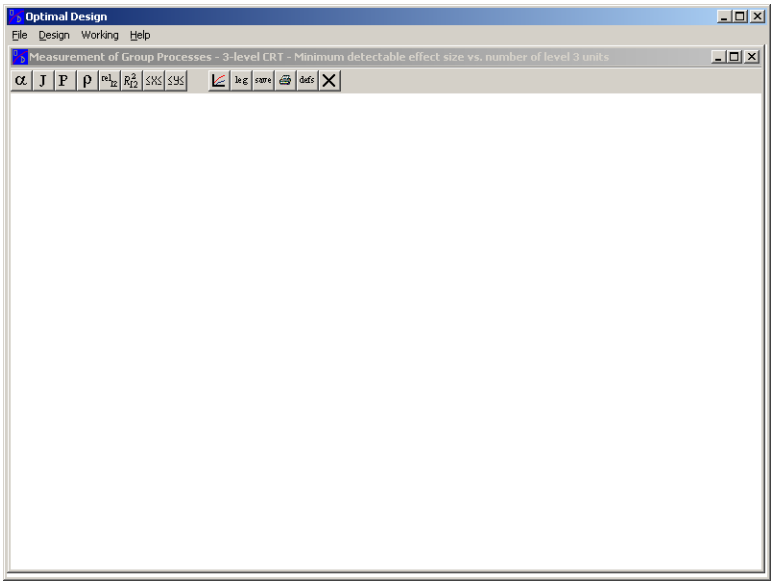


Figure 19.3. Blank screen for MDES vs. number of clusters.

Step 2: Click on *J*. Set *J*(1) = 10.

Step 3: Click on *P*. Set *P*(1) = .80.

Step 4: Click on ρ. Set rho (1) = 0.15

Step 5: Click on relL2. Set Reliab(l2)(2) = 0.75. The resulting power curve appears in Figure 19.4.

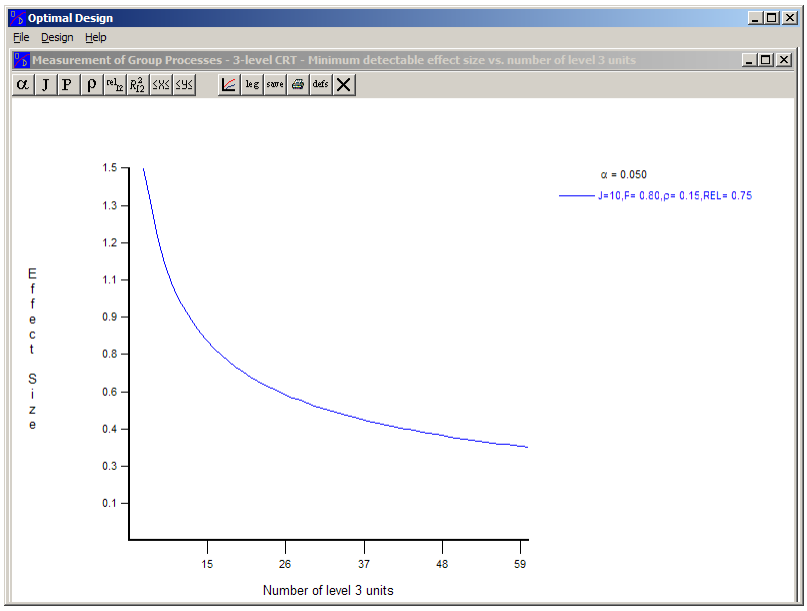


Figure 19.4. Power vs. number of level 3 units.

Clicking on the trajectory reveals that with 40 clusters, the MDES is approximately 0.47.

The examples in this section are meant to provide at guide to users for how to use the 2-level CRT with a cluster level outcome. We described Power vs. number of level-3 units (*K*) and MDES vs. number of level-3 units (*K*). The other options function similarly, and simply place a different parameter on the x-axis. The choice of which module is most appropriate depends on the unknown parameters. However, all modules yield the same results if identical parameters are used so the choice depends on what module is most closely aligned with the known and unknown parameters in a study.

20.0 Multi-site cluster randomized trials with a cluster-level outcome

A multi-site cluster randomized trial with a cluster level outcome is a blocked design. For example, imagine an evaluation for a new character development program. Within schools, classrooms are randomly assigned to either the new program or their regular program. The effects of adopting the new program are evaluated by the quality of functioning of a classroom. This means that the outcome is being measured at the level two/cluster level. The measurement usually relies on either observational or interview data, which means it includes measurement error. The measurement error as well as the nested structure of the data has implications for statistical power. The treatment of blocks, or schools, as fixed or random effects also has implications for power. First, we take a closer look at the models.

20.1The model

We can represent data from a MSCRT as a three level model where the lowest level is the measurement model, level-two is the cluster level model, and the highest level is the site/block level model. Note that the cluster is the unit of randomization and the unit of measure. Any variation below the level of the cluster contributes to the measurement error including temporal variation, observer variation, individual variation, or item variation. Assuming items constitute the sole source of error, the level-1 model is:

$$Y_{tjk} = \pi_{0jk} + e_{tjk} \qquad e_{tjk} \sim N(0, \sigma^2) \qquad [20.1]$$

for

$t = 1,..., T$ items within a scale

$j = 1,..., J$ clusters

$k = 1,..., K$ blocks

where

Y_{tjk} is the observed outcome for item t in cluster j in block k

π_{0jk} is the mean for cluster j in block k

e_{tjk} is random error associated with the item

Note that equation 1 represents a very simple measurement model and temporal or observer variation could also be modeled at this level.

The level-2 model is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} W_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi) \qquad [20.2]$$

where

β_{00k} is the mean for site *k*

β_{01k} is the treatment effect at site *k*

W_{jk} is a treatment contrast indicator, ½ for treatment and -½ for control

r_{0jk} is random error associated with the cluster

The level-3 model is

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}}) \; u_{01k} \sim N(0, \tau_{\beta_{11}}) \\ \beta_{01k} &= \gamma_{010} + u_{01k} \end{aligned} \qquad [20.3]$$

$$\text{cov}(u_{00k}, u_{01k}) = \tau_{\beta_{01}}$$

where

γ_{000} is the grand mean

γ_{010} is the average treatment effect

u_{00k} is the random error associated with each site mean

u_{01k} is the random error associated with each site treatment effect.

20.2 Testing the treatment effect

We are interested in the main effect of treatment, γ_{010}, which is estimated by:

$$\hat{\gamma}_{010} = \bar{Y}_E - \bar{Y}_C \qquad [20.4]$$

where $\bar{Y}_E$ is the mean for the experimental group and $\bar{Y}_C$ is the mean for the control group. Assuming balanced allocation of clusters to treatment and control, the variance of the estimated treatment effect is:

$$Var(\hat{\gamma}_{010}) = \frac{\tau_{\beta_1} + 4\left(\frac{\tau_\pi + \sigma^2 / T}{J}\right)}{K} \qquad [20.5]$$

where *T* is the number of items within the scale, *J* is the number of clusters per site, and *K* is the total number of sites.

To test the null hypothesis of no treatment effect, we use an *F* statistic.

$H_0 : \gamma_{010} = 0$ vs. $H_1 : \gamma_{010} \neq 0$,

The *F* test follows a non-central *F* distribution, $F(1, K\text{-}1; \lambda)$. Recall that the noncentrality parameter, λ, is a ratio of the squared-treatment effect to the variance of the treatment effect estimate. Below is the noncentrality parameter for the test.

$$\lambda = \frac{\gamma_{010}^2}{Var(\hat{\gamma}_{010})} = \frac{K\gamma_{001}^2}{\tau_{\beta_1} + 4\left(\frac{\tau_\pi + \sigma^2/T}{J}\right)}.$$

20.3 Standardized Notation

To reinforce the concept of the measurement model, we standardize the model differently and reconceptualize it in terms of reliability instead of intra-class correlations. We define the standardized effect size below:

$$\delta = \frac{\gamma_{010}}{\sqrt{\tau_\pi}} \qquad [20.6]$$

Note that the effect size is divided by the square root of the between cluster variance. Recall that this is because we are standardizing at the same level that we are randomizing. We must also define the effect size variability in this same metric as the effect size. It is:

$$\sigma_\delta^2 = \frac{\tau_{\beta_1}}{\tau_\pi} \qquad [20.7]$$

In addition to the new effect size and effect size variability, we define the cluster level reliability as:

$$reliab_{L2} = \frac{\tau_\pi}{\tau_\pi + \sigma^2/T} \qquad [20.8]$$

We want the level-2 reliability because we are randomizing and measuring at level-2[7].

The non-centrality parameter can be defined in terms of the new effect size and the level-2 reliability. The non-centrality parameter is:

[7] Note that the standardized values depend on the user knowing the within-block between-cluster variance, τ_π. Because this value is typically unknown prior to blocking, the Optimal Design program asks for the standardized values prior to blocking as well as the percentage of variance explained by the blocking variable and calculates the standardized values within the program.

$$\lambda = \frac{K\delta^2}{\sigma_\delta^2 + 4/J(reliab_{L2})} \qquad [20.9]$$

Recall that the larger the non-centrality parameter, the greater the power of the test. Thus it is clear that increasing the reliability, the number of sites, and the number of clusters per site are three options that the researcher has to increase the power. Increasing the effect size also increases power but is typically not set by the researcher.

20.4 Using the Optimal Design for multisite cluster randomized trials with a cluster-level covariate

This section focuses on how to use the Optimal Design software to design a multisite cluster randomized trial with a cluster-level outcome. The menu for the MSCRT with a cluster level covariate is shown below and can be found by clicking on the following: Design → Cluster randomized trials with cluster-level outcomes (measurement of group processes) → Multisite (or blocked) cluster randomized trials → Treatment at level 2. The menu is below.

Power on y-axis

- Power vs. number of sites (*K*)
- Power vs. number of clusters (*J*)
- Power vs. effect size (delta)
- Power vs. cluster-level reliability (Rel(L2))

MDES on y-axis

- MDES vs. number of sites (*K*)
- MDES vs. number of clusters (*J*)
- MDES vs. power (*P*)
- MDES vs. cluster-level reliability (Rel(L2))

The first set of options present the power on the y-axis and either the number of sites, number of clusters, effect size, or the cluster-level reliability on the x-axis. The second set of options present the MDES on the y-axis and either the number of sites, number of clusters, power, or cluster-level reliability on the x-axis. We present an example below and guide the user through the steps for approaching the example via the power determination approach or the effect size approach.

20.5 Example

Suppose a team of researchers want to measure the impact of an intervention on the quality of functioning in elementary classrooms. The researchers hypothesize that the overall quality of functioning will increase with participation in their program. They plan to assess the quality of the classrooms using an observational instrument. The researchers plan to randomly assign 8 classrooms within schools to either the treatment or control condition. Thus schools act as blocks, or sites, and classrooms are within the sites. They hypothesize that blocking will explain 20% of the variation in the outcome. The researchers expect there to be variability across sites so they plan to use a random effects model. Based on previous studies, they expect the effect size variability to be approximately 0.10. Section 20.6 presents a scenario in which the power determination approach for conducting a power analysis is most applicable and the details of how to do the power analysis using OD. Section 20.7 presents a scenario in which the effect size approach is most applicable the details of how to do the power analysis using OD.

20.6 Power determination approach for conducting a power analysis

Based on previous studies that use the same observational instrument, the researchers estimate the reliability equals 0.75 and want to be able to detect a minimum effect size of 0.50. How many schools are required in order to achieve power equal to 0.80?

In Scenario 1, the number of schools or sites, *K*, is unknown. As a result, we want to select the power vs. number of clusters (*K*) option. This allows the number of sites to vary along the x-axis. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with cluster level outcomes (measurement of group processes) → Multisite cluster randomized trials → Treatment at level 2 → Power on y-axis → Power vs. number of sites (*K*) as shown in Figure 20.1.

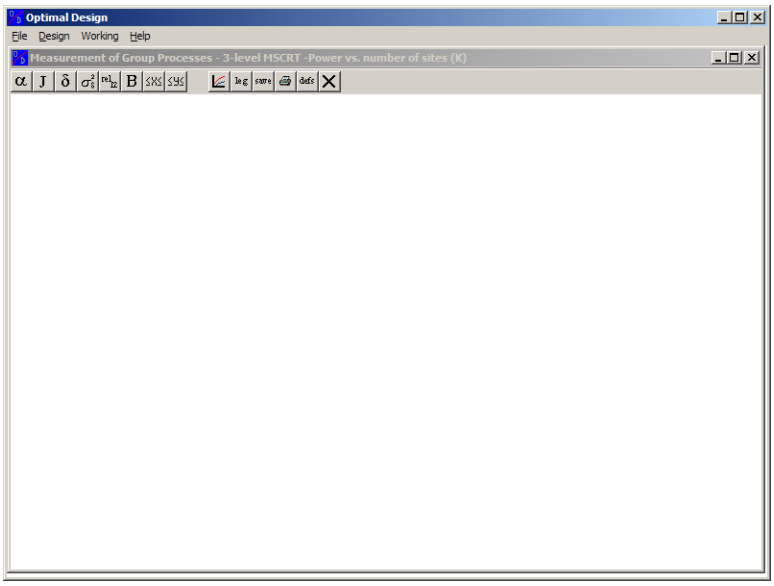


Figure 20.1. Blank screen for Power vs. number of sites (*K*).

The toolbar at the top includes the parameters required for calculating the power: the number of clusters per site (*J*), the effect size (δ), the effect size variability (σ_δ^2),the reliability of the cluster-level covariate (relL2), and the percent of variance explained by blocking. The number of sites (*K*) does not appear on the toolbar because it varies along the x-axis.

Step 2: Click on *J*. Set *J*(1) = 8.

Step 3: Click on δ. Set delta(1) = 0.50. Recall that delta in the case of a 2-level CRT for a cluster level covariate is defined as $\delta = \frac{\gamma}{\sqrt{\tau}}$. The default power curves appear. However, we must first set the additional parameters to match the values in the particular example before we interpret the curves.

Step 4: Click on σ_δ^2. Set σ_δ^2=0.10.

Step 5: Click on relL2. Set relL2 (1) = 0.75.

Step 6: Click on B. Set B(1) = 0.20. The resulting curve appears in Figure 20.2.

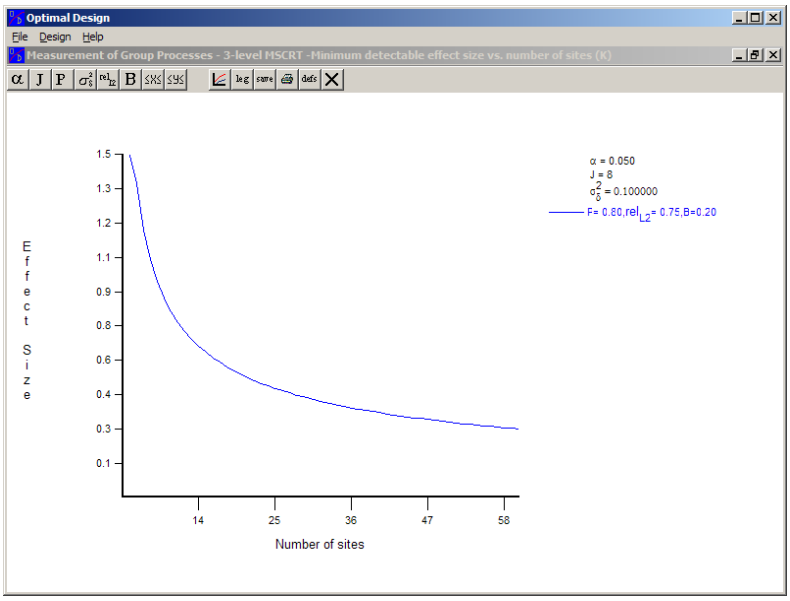


Figure 20.2. Power curve.

Clicking along the power curve, we can see that approximately 23 total sites or schools are required for the study with 8 classrooms per school.

The example provided in this section placed the sample size on the x-axis. However, the reliability or the effect size could be placed on the x-axis and the steps could easily be adapted to conduct the power analysis.

20.7 Effect size approach for conducting a power analysis

Based on previous studies that use the same observational instrument, the researchers estimate the reliability equals 0.75. They have secured 8 classrooms per school and 20 schools. What is the MDES with power = 0.80?

In Scenario 2, the MDES is unknown so it makes more sense to select an option with the MDES on the y-axis. One option is to select MDES vs. number of sites (*K*). This will allow the user to see how the MDES changes as a function of the number of sites holding all other parameters constant. Using this approach is very useful but also requires that after the MDES is determined, the researcher consult the literature or findings from a pilot study to determine if the MDES is reasonable. The steps for conducting the power analysis follow.

Step 1: Select Design → Cluster randomized trials with cluster level outcomes (measurement for group processes) -> Multisite cluster randomized trials → Treatment at level 2 → MDES on y-axis → MDES vs. number of sites (*K*) as shown in Figure 20.3.

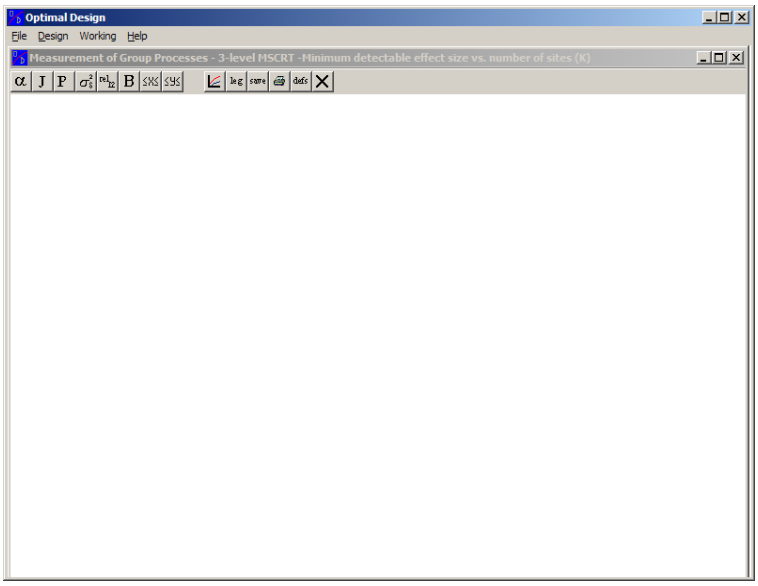


Figure 20.3. Initial blank screen for MDES vs. number of sites (*K*).

Step 2: Click on *J*. Set *J*(1) = 8.

Step 3: Click on *P*. Set *P*(1) = 0.80.

Step 4: Click on σ_δ^2. Set σ_δ^2=0.10.

Step 5: Click on relL2. Set Reliab(l2)(1) = 0.75.

Step 6: Click on B. Set B(1) = 0.20. The resulting power curve appears in Figure 20.4.

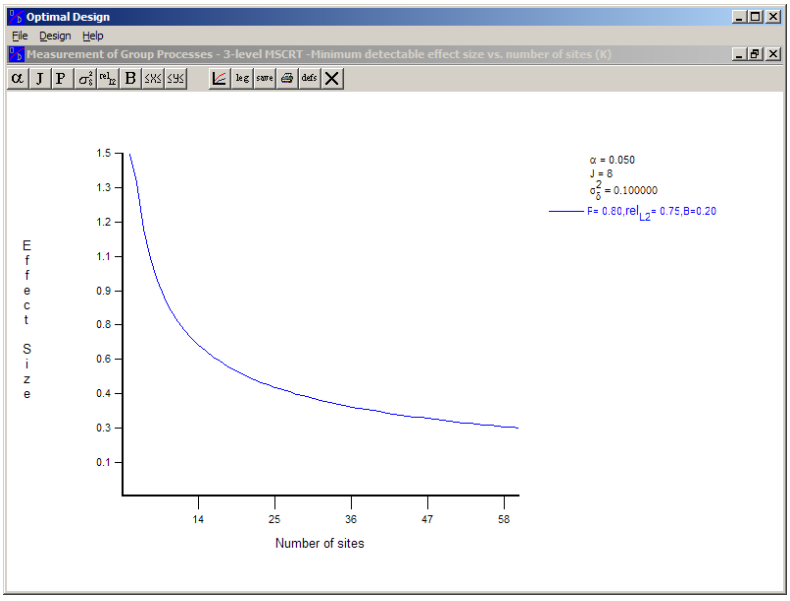


Figure 20.4. Power curve.

Clicking along the trajectory reveals a MDES of 0.55 with 20 schools.

The examples in this section are meant to provide at guide to users for how to use the MSCRT with a cluster level outcome. We described Power vs. number of clusters (*J*) and MDES vs. Power (*P*). The other options function similarly, and simply place a different parameter on the x-axis. The choice of which module is most appropriate depends on the unknown parameters. However, all modules yield the same results if identical parameters are used so the choice depends on what module is most closely aligned with the known and unknown parameters in a study.

APPENDIX A.1: Background Information on the Elementary, Middle, and High School Reading and Math Data

Data from five large urban school districts were used to obtain empirical estimates of variance components and corresponding ICCs and R-square values (described below). In these settings, students were clustered within schools, and schools were blocked by district. These data were analyzed using a 3 level multisite cluster randomized model (3 level MSCRT; see section 7.0 in the documentation) with districts treated as fixed effects.

Reading and math test scores are available for grades 3, 5, 8, and 10 from each district. For each of these test scores, the empirical estimates module for the 3 level MSCRT shows:

(1) the unconditional intraclass correlation (ICC), that is, the proportion of total student variation in an outcome within a district that is across schools;
(2) the proportion of outcome variation across students within schools that is explained (R-square) by a specified set of student level or school-level baseline characteristics (described below); and
(3) the proportion of outcome variation across schools that is explained (R-square) by a set of student-level and/or school-level baseline characteristics.

As noted in Section 9.1, we can represent data from a multi-site cluster randomized trial as a three level model, with students nested within schools nested within districts.

The level-1 model, or student-level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \qquad [9.1]$$

for $i \in \{1,2,...,n\}$ students per school, $j \in \{1,2,...,J\}$ schools and $k \in \{1,2,...,K\}$ districts,

where π_{0jk} is the mean for school *j* in district *k*;

e_{ijk} is the error associated with each student; and

σ^2 is the within-school variance.

The level-2 model, or cluster-level model, is:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi) \qquad [9.2]$$

where β_{00k} is the mean for district *k*;

r_{0jk} is the random effect associated with each school; and

τ_π is the variance between schools within districts.

The level-3 model, or district-level model, specifies district fixed effects:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad [9.15]$$

where γ_{000} is the grand mean;

u_{00k}, for $k \in \{1,2,...,K\}$, are fixed effects associated with each district mean, constrained to have a mean of zero.

A version of the above model was estimated for each test score outcome in each grade within each of the five districts. Estimated variances from the unconditional models were used to obtain the ICCs.

Next, the models were estimated with a number of different combinations of student-level covariates in equation 9.1 and school-level covariates in equation 9.2:

Covariate (s)	**Description / Definition**
Y_{-1}	Mean school score for the same test in the same grade in the previous year (i.e., lagged one year)
Y_{-2}	Mean school score for the same test in the same grade two years previously (i.e., lagged two years)
Y_{-1}, Y_{-2}	Mean school score for the same test in the same grade, lagged one year & lagged two years
y_{-1}	Individual score for the same test in a previous year (i.e., lagged one year)
y_{-2}	Individual score for the same test two years previously (i.e., lagged two years)
y_{-1}, y_{-2}	Individual score for the same test, lagged one year & lagged two years
Y_{-1}, y_{-1}	Mean school score for the same test in the same grade lagged one year & individual score for the same test lagged one year
Z_{-1}	Mean school score for a different test (i.e., reading instead of math or vice versa) in the same grade in the previous year (i.e., lagged one year)
z_{-1}	Individual score for a different test (i.e., reading instead of math or vice versa) in the previous year (i.e., lagged one year)
X	Vector of student demographic characteristics (varies somewhat across schools, but typically includes gender, race/ethnicity, free/reduced lunch)
X, Y_{-1}	Vector of student demographic characteristics & mean school score for the same test in the same grade lagged one years
X, y_{-1}	Vector of student demographic characteristics & individual score for the same test, lagged one year

R-square values at the student level and at the school level were generated based on models that incorporate each of the above covariates or covariate sets.

Data Source:

Bloom, Howard S., Lashawn Richburg-Hayes, and Alison Rebeck Black. 2005. “Using

Covariates to Improve Precision: Empirical Guidance for Studies That Randomize Schools to Measure the Impacts of Educational Interventions," MDRC Working Papers on Research Methodology.
http://www.mdrc.org/publications/417/abstract.html

The following summary of the data is quoted directly from the working paper, p. 18:

> The present empirical analysis is based on individual data for thousands of students from hundreds of schools located in five urban school districts. Elementary school analyses focus on reading and math test scores in grades three and five using data from all five districts. Middle school analyses focus on reading and math test scores in grade eight and the high school analyses focus on reading and math test scores in grade 10. Data for middle school and high school analyses were only available for two of the five districts. All analyses were also replicated for as many years as possible in each district.
>
> Table 3 [shown on p. 47 of the working paper] briefly describes the districts, schools, and students in the sample for the present analysis. First note that the districts in the sample are fairly large. They represent from 25 to 168 elementary schools, 17 to 41 middle schools, and 11to 32 high schools. The average elementary school in each district had 57 to 75 third-grade students who were tested in a given year; the average middle school had 196 to 297 eighth-grade students; and the average high school had 234 to 269 tenth-grade students. In two districts, students were predominantly black; in two other districts they were a mix of blacks and Hispanics; and in the fifth district information was not available on their background characteristics. In the three districts where data on economic status were available for elementary schools, the percentage of students who were categorized as low-income ranged from 41 percent to 79 percent.

APPENDIX A.2: Pre-K: Social-emotional and Cognitive

Five studies of early childhood programs were used to obtain empirical estimates of variance components and corresponding ICCs and R-square values (described below). Generally for these studies, children are clustered within classrooms, classrooms are clustered in centers, with centers blocked by grantee.[8] The data were analyzed using a 4 level multisite cluster randomized model (4 level MSCRT; see section 10.0 of the documentation), with fixed effects at the fourth, or grantee, level to obtain the ICCs and R-square values.

A large number of measures or indicators are available from each study. For each measure, the empirically based output for an MDES for the 4 level MSCRT include:

(1) unconditional intraclass correlations (ICCs) for centers and classrooms that represent the proportion of student variation that exists within each level of clustering;
(2) an R-square value for students, classrooms and centers that represents the proportion of outcome variation at each of these levels that is explained by a specified set of covariates.

As noted in Section 10, a four-level multi-site cluster randomized trial can be written as a four level hierarchical model. The level-1 model for children is:

$$Y_{ijkl} = \pi_{0\,jkl} + e_{ijkl} \qquad e_{ijkl} \sim N(0, \sigma^2) \qquad [10.1]$$

for $i \in \{1,2,...,n\}$ children per classroom, $j \in \{1,2,...,J\}$ classrooms per center, $k \in \{1,2,...,K\}$ centers per grantee, and $l \in \{1,2,...,L\}$ grantees,

where $\pi_{0\,jkl}$ is the mean for classroom *j* in center *k* in grantee *l;*

e_{ijkl} is the error associated with each child; and

[8] An exception is the 4Rs study, which examined children in 3rd grade classrooms. These data are more appropriately viewed as a 3 level MSCRT, like the ones described in Appendix B.

σ^2 is the within-classroom variance.

The level-2 model for classrooms within centers is:

$$\pi_{0jkl} = \beta_{00kl} + r_{0jkl} \qquad r_{0jkl} \sim N(0, \tau_\pi) \qquad [10.2]$$

where β_{00kl} is the mean for center *k* in grantee *l;*

r_{0jkl} is the random effect associated with each classroom; and

τ_π is the variance between classrooms within centers.

The level-3 model for centers within grantees is:

$$\beta_{00kl} = \gamma_{000l} + u_{00kl} \qquad u_{00kl} \sim N(0, \tau_\beta) \qquad [10.3]$$

where γ_{000l} is the mean for grantee l;

u_{00kl} is the random effect associated with center; and

τ_β is the variance between centers within grantees.

The level four model for grantees is estimated as a fixed effects model at level 4:

$$\gamma_{000l} = \eta_{0000} + s_{000l} \qquad [10.14]$$

where η_{0000} is the grand mean;

s_{000l}, are fixed effects associated with each grantee mean, constrained to have a mean of zero.

A version of the above 4-level model was estimated for each measure in each dataset described below. Estimated variance components from the unconditional models were used to obtain the ICCs.

Next, the models were estimated with child-level covariates in equation 10.1 and classroom-level covariates in equation 10.2. The child-level covariates varied somewhat from study to study, but included typical demographics (e.g. child's gender, race/ethnicity, age, family income, mother's education, mother's age, mother's marital status, mother's

employment status) as well as a baseline measure (from the fall of the same school year as the outcome) for the outcome of interest. The classroom-level covariates also varied somewhat, but typically included class size, age of teacher, whether the teacher had a B.A. or Child Development Associate (CDA) certificate, and years of teaching experience[9].

A set of R-square values are available in the current version of the empirical estimates module, obtained from models that include a set of child-level demographic characteristics, a baseline measure of the same outcome, and a set of classroom-level characteristics. Note that the R-square values range from 0 to 1.0 at each level across the various datasets.

Data Sources:

FACES 1997 & 2003

Study background: The Family and Child Experiences Survey (FACES) gathers longitudinal data on the cognitive, social, emotional, and physical development of Head Start children; the characteristics, well-being, and accomplishments of their families; the quality of Head Start classrooms; and the characteristics, needs, and opinions of Head Start teachers and other program staff. Four cohorts of FACES—each a nationally-representative sample—have been fielded to date – FACES 1997, 2000, 2003, and 2006. At the time the information for this project was generated, access was available to FACES 1997, 2000, and 2003. The sampling strategy for FACES 2000 made it unsuitable for providing the information needed and hence it was not used.

1997 National Cohort

Design: FACES 1997 used a three-stage sampling procedure. The first stage selected 40 Head Start grantees from among the 1,734 grantees nationwide at the time. This selection was organized by 16 strata based on region (Northeast, Midwest, South, and West), Metropolitan Statistical Area (MSA) status (urban, rural), and percent minority

[9] Treatment indicators were also included in the models for the experimental studies described below (CSRP, REDI, 4Rs).

enrollment (above 50%, below 50%). The second stage selected 180 centers from the 40 Head Start grantees. The third stage selected roughly 3,200 students from the 180 centers. Participating 3- and 4-year old children were assessed in Fall 1997, Spring 1998, and Spring 1999.

Sample: Baseline information was obtained from 3,006 of the students in the sample. This information indicates that: 43 percent of the students lived with both of their parents, seventy-two percent of the mothers of students had at least a high school diploma or GED, less than 9 percent of these mothers had an Associate's degree or higher, forty-two percent of households reported less than $1,000 in monthly income (from all sources), 85 percent of households received supplemental income from one or more sources, one-fifth of children were reported to have been exposed to community or domestic violence.

2003 National Cohort

Design and Sample: A stratified four-stage sampling design, similar to that for FACES 1997 was used to draw the sample for FACES 2003. 68 Head Start grantees were selected in the first stage of this process. In the second stage, 175 centers (aggregated into 110 "center groups") were selected. In the third stage, 409 classes were selected from the 110 center groups. In the fourth stage 2,816 children were chosen from the 409 classes. The study collected data in Fall 2003, Spring 2004, and Spring 2005.

For further information:

Study: http://www.acf.hhs.gov/programs/opre/hs/faces/

Measures:

http://researchconnections.org/childcare/resources/18961?q=faces+instrument+matrix

Chicago School Readiness Project (CSRP)

Intervention: The Chicago School Readiness Project intervention model comprised four specific components: (1) teacher training in behavior management strategies, (2) "coaching" to help teachers implement these strategies, (3) stress reduction workshops, and (4) direct services for children with severe emotional or behavioral problems.

Design: The intervention was implemented for two cohorts of children and teachers, with Cohort 1 participating from fall to spring in 2004-05 and Cohort 2 participating from fall to spring in 2005-06. A total of 18 Head Start centers were randomized in matched pairs to the intervention or a control condition.[10] Children in the participating centers were assessed for at least two time points during the Head Start year (baseline and follow-up 1 in May), and were tracked into Kindergarten (follow-up 2) and 1st grade (follow-up 3).

Sample: Sites were selected on the basis of (a) receipt of Head Start funding, (b) having two or more classrooms with "full day" programming, and (c) being located in one of seven selected high-poverty neighborhoods in Chicago. Eighteen sites from the seven neighborhoods were included and two classrooms from each site were included for a total of 35 classrooms (one classroom dropped out). Sites were then matched into pairs based on their demographic characteristics and randomized within pairs to the intervention or the control group. At baseline, 543 children participated in the CSRP study with 509 remaining by the spring follow-up. The majority of the sample is African American (65%), followed by Hispanic (28%), and 69% of the families are headed by a single parent. To estimate the underlying ICCs and R-squares from this data set the fact that sites were matched into pair was ignored.

For further information:

Study: http://steinhardt.nyu.edu/ihdsc/csrp/

Head Start REDI (Three Pennsylvania Counties)

[10] Indicators for matched pairs were *not* included in the models we used to estimate variance components for this program.

Intervention: The Preschool PATHS Curriculum was used to promote children's social-emotional skills. It targets four domains: (1) pro-social friendship skills, (2) emotional understanding and emotional expression skills, (3) self-control, and (4) problem-solving skills, including interpersonal negotiation and conflict resolution. In addition, four language and emergent literacy skills were targeted: (1) vocabulary, (2) syntax, (3) phonological awareness, and 4) print awareness.

Design: Centers were stratified based on their location, length of grantee, and student demographics. Within each stratum centers were randomly assigned to the intervention or a control group. Most classrooms (67%) were in small centers (1-2 classrooms), but 4 centers were larger, containing 3-5 classrooms. Two cohorts were studied, in 2004-05 and 2005-06. To estimate the underlying ICCs and R-squares from this data set the fact that centers were stratified was ignored.

Sample: Participants included two cohorts of 4-year-old children (Total N=356; 17% Hispanic, 25% African American; 54% girls) in 44 Head Start classrooms in three counties in Pennsylvania. Half of the classrooms came from a large county, which included an urban community surrounded by smaller communities. The other classrooms came from smaller counties characterized by small towns and rural areas. Children were recruited across two years from non-overlapping classrooms.

For further information:

Study: http://headstartredi.ssri.psu.edu/

Measures: http://headstartredi.ssri.psu.edu/measures_public

Reading, Writing, Respect and Resolution, 4Rs (New York City)

Intervention: A school-based program that integrates social and emotional learning into the language arts curriculum for grades K-5. The 4Rs uses high quality children's literature to help students gain skills and understanding in several areas including

handling anger, listening, cooperation, assertiveness, and negotiation. It provides a pedagogical link between teaching conflict resolution and fundamental academic skills. The intervention was implemented school-wide for 3 consecutive school years.

Design: Schools were matched into pairs and then randomized within pairs to the intervention or a control group[11]. A cohort of 3rd grade children, their teachers, and their classrooms were tracked 5 times post-baseline in Fall and Spring of each school year through the end of 5th grade (2004-05, 2005-06, and 2006-07). (Note: The variance components used to calculate MDES from this study were from the 2004-05 year.) To estimate the underlying ICCs and R-squares from this data set the fact that schools were matched into pairs was ignored.

Sample: 18 New York City elementary schools were recruited, matched into pairs, and randomized to the intervention or a control group. In the first year, participants included 942 children (51.3% female) and 85 teachers/classrooms (94% female). At baseline, child race/ethnicity was 48% Hispanic/Latino, 38% Black/African American, 5% White, non-Hispanic, and 8% other.

For further information:

Aber, J. Lawrence; Brown, Joshua L.; Jones, Stephanie M.; Berg, Juliette; and Torrente, Catalina, "School-based strategies to prevent violence, trauma, and psychopathology: The challenges of going to scale" (2011). *Psychology Faculty Publications.* Paper 132. http://fordham.bepress.com/psych_facultypubs/132

Jones, Stephanie M.; Brown, Joshua L.; and Aber, J. Lawrence, "Two-Year Impacts of a Universal School-Based Social-Emotional and Literacy Intervention: An

[11] Indicators for matched pairs were *not* included in the models we used to estimate variance components for this program.

Experiment in Translational Developmental Research" (2011). *Psychology Faculty Publications.* Paper 131. http://fordham.bepress.com/psych_facultypubs/131

Brown, J. L., Jones, S. M., LaRusso, M. D., & Aber, J. L. (2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs Program. Journal of Educational Psychology, 102(1), 153-167.

Jones, Stephanie M.; Brown, Joshua L.; Hoglund, Wendy L. G.; and Aber, J. Lawrence, "A School-Randomized Clinical Trial of an Integrated Social–Emotional Learning and Literacy Intervention: Impacts After 1 School Year" (2010). *Psychology Faculty Publications.* Paper 130. http://fordham.bepress.com/psych_facultypubs/130

Appendix B: Meta-Analysis

Model specification and estimation, hypothesis test and power computations

Given a number of studies sharing the same hypothesis, the idea is to obtain the common minimum detectable effect size at a given level of statistical power based on select summary statistics from each of the studies. In other words, the goal is to perform a power analysis for a meta-analysis (also called a research synthesis).

Our formulation rests on the following assumptions:

1. The studies to be included in the meta-analysis share a common hypothesis.
2. Estimates of the effect size and the standard error of the effect size are accessible in standardized form for each study and have been correctly calculated.
3. The studies are independent.

The data available can be seen as having a hierarchical structure in which the units under study are nested within studies, with some variation lying between the units in any one particular study and some variation lying across studies. A basic task the meta-analyst faces is to determine whether the results are consistent across studies and, if the results are not consistent, determine why the results change from any one study to the next. Hierarchical models provide a helpful framework for dealing with this type of data (Raudenbush and Bryk, 2002, Chapter 7).

In a meta-analysis, typically summary statistics are available from each study while the raw data are rarely accessible. The statistical model specified in this section is built on this premise. That is, we assume estimates of the treatment effect size and corresponding standard error are accessible from each study.

One potential complication is that outcome measures are not necessarily the same across individual studies, even if they target the same concept. To address this problem, meta-analysts typically use standardized measures of treatment effects, translating the results of any one particular study to a common scale. Suppose there are J studies being considered, each study having a standardized effect size estimate denoted by d_j, for $j = \{1,2,,...,J\}$. In the context of program evaluation, this standardized effect size might be the standardized mean difference between the experimental and the control groups, given by

$$d_j = \frac{\bar{Y}_j^E - \bar{Y}_j^C}{S_j} \qquad \text{[B.1]}$$

Where $\bar{Y}_j^E$ is the average for those in the experimental condition for study j;

$\bar{Y}_j^C$ is the average for those in the control condition for study j; and

S_j is the pooled standard deviation for study *j*.

In other words, d_j is an estimate of the mean difference between the experimental and control groups in standard deviation units for study *j*.

Suppose estimates of the treatment effect size and of the standard error of the treatment effect size are available for each of *J* studies. Data can then be modeled with a two-level hierarchical model. The level-1 (within-studies) model would be

$$d_j = \delta_j + e_j, \text{ with, } e_j \sim N(0, V_j) \qquad \text{[B.2]}$$

where δ_j is the treatment effect for study *j* and

e_j is the level-1 random error, normally distributed with mean 0 and known variance V_j.

The unconditional level-2 (between-studies) model would be

$$\delta_j = \theta + u_j \qquad \text{with } u_j \sim N(0, \tau), \qquad \text{[B.3]}$$

where θ is the average treatment effect across studies (true effect size) and

u_j is a level-2 random error, normally distributed with mean 0 and variance τ.

The level-1 and level-2 models above yield the following mixed model:

$$d_j = \theta + u_j + e_j, \qquad \text{with } e_j \sim N(0, V_j) \text{ and } u_j \sim N(0, \tau). \qquad \text{[B.4]}$$

From the model above, it follows that $E(d_j) = \theta$ and

$$\begin{aligned} \text{Var}(d_j) &= \text{Var}(u_j) + \text{Var}(e_j) \\ &= \tau + V_j \\ &= \text{parameter variance} + \text{error variance} \\ &= \Delta_j. \end{aligned}$$

If sample sizes vary across studies, each d_j can be seen as an unbiased estimator of θ with variance Δ_j. The precision of d_j is defined as the reciprocal of its variance:

$$\text{Precision}(d_j) = \Delta_j^{-1}.$$

If the Δ_j's are known (or estimated consistently, the maximum likelihood estimator (minimum-variance unbiased estimator) of θ is the precision-weighted average:

$$\hat{\theta} = \frac{\sum_{j=1}^{J} \Delta_j^{-1} * d_j}{\sum_{j=1}^{J} \Delta_j^{-1}}. \qquad [B.5]$$

If all the studies have the same sample size, the expression above can be reduced to the simple average

$$\hat{\theta} = \frac{\sum_{j=1}^{J} d_j}{J}. \qquad [B.6]$$

The precision of θ is the sum of its precisions:

$$\begin{aligned} \text{Precision}\left(\hat{\theta}\right) &= \sum_{j=1}^{J} \Delta_j^{-1} \\ &= \sum_{j=1}^{J} \left(\tau + V_j\right)^{-1} \end{aligned} \qquad [B.7]$$

and its variance is the inverse of its precision

$$\text{Var}(\hat{\theta}) = \left[\sum_{j=1}^{J} \left(\tau + V_j\right)^{-1}\right]^{-1}. \qquad [B.8]$$

Example

Suppose a researcher is interested in calculating the power of a meta-analysis with 19 studies. For each individual study, the researcher has an estimate of the effect size and its variance. The data is displayed in Table B.1.

Table A.1. Sample data for meta-analysis. Column 1 corresponds to the effect size for the study and column 2 corresponds to the variance.

0.03	0.016
0.12	0.022
-0.14	0.028
1.18	0.139
0.26	0.136
-0.06	0.011
-0.02	0.011
-0.32	0.048
0.27	0.027
0.8	0.063
0.54	0.091
0.18	0.050
-0.02	0.084
0.23	0.084
-0.18	0.025
-0.06	0.028
0.3	0.019
0.07	0.009
-0.07	0.030

Note. This data is from the effect of teacher expectancy on Pupil IQ (from Chapter 7 of Raudenbush and Bryk, 2002, p. 211).

The following steps are needed to estimate the power of a meta-analysis:

Step1: Click on Design -> Meta-analysis -> Read Data/Generate Variance. Figure B.1 displays the menu that appears.

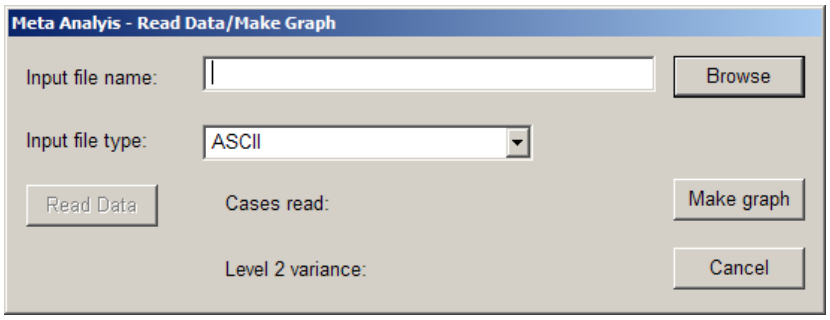


Figure B.1. Menu for meta-analysis.

Step 2: Click on Browse. Find the file that contains the data. The data file can be in SPSS, Excel, or any package. However, it should be saved as a .dat or a .csv file. The data needs to be organized in two columns, the first with the effect size and the second with the variance. The number of rows should correspond to the number of studies being considered. Table B.1contains sample data for this example.

Step 3: Click on Read Data. This allows the OD to read the data. The number of cases read and level 2 variance will be displayed as shown in Figure B.2.

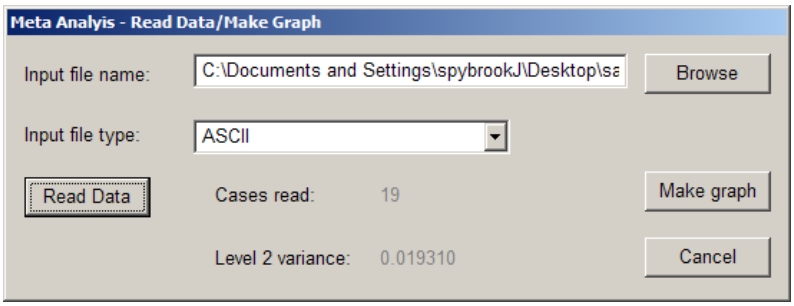


Figure B.2. Cases read and level two variance.

Step 4: Click on Make Graph. The graph appears in Figure B.3. Clicking on the graph reveals power of 0.80 for an effect size of 0.144.

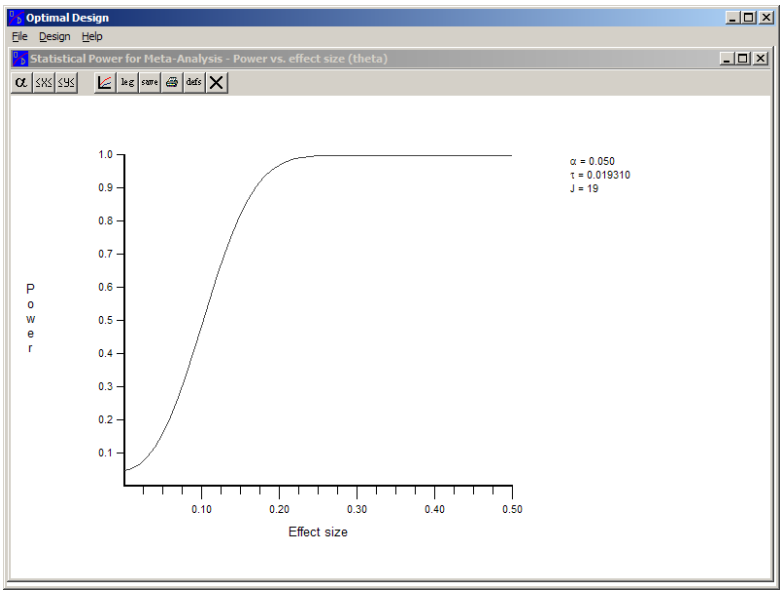


Figure B.1. Power curve for meta-analysis.

Appendix C – Optimal Sample Allocation

The cluster randomized trial with person level outcomes, treatment at level 2, also has an option for optimal sample allocation.

The total variable cost of data collection can often be reasonably approximated by the formula below:

$$T = J(C_1 n + C_2) \qquad \text{[C.1]}$$

where J = number of clusters;

n = number of participants within a cluster;

C_1 = cost per participant;

C_2 = cost per cluster; and

T = total cost.

To calculate the optimal sample size, first find the optimal n and then find the optimal J. The optimal n in this case is that which minimizes the variance of the treatment effect in equation C.2:

$$Var(\hat{\gamma}_{01}) = \frac{4(\tau + \sigma^2 / n)}{J}. \qquad \text{[C.2]}$$

Substituting $J = \frac{T}{nC_1 + C_2}$ (a simple rearrangement of the cost equation) and minimizing the equation with respect to n, we obtain the formula for optimal n:

$$n_{opt} = \frac{\sigma}{\sqrt{\tau}} * \sqrt{\frac{C_2}{C_1}} \qquad \text{[C.3]}$$

where σ is the within cluster standard deviation;

$\sqrt{\tau}$ is the between cluster standard deviation;

C_1 is the cost per person; and

C_2 is the cost per cluster

From the formula, we can see as the within-cluster variance increases relative to the between-cluster variance, optimal n increases. Intuitively this makes sense. If there is large variation within clusters, we would want to sample more people in each cluster to represent that variation. However, if the within cluster variation is very small, optimal n decreases. In this case,

we want fewer people in each cluster because most of the variation is between clusters so adding more people will not be very helpful. In terms of the cost ratio, if the cost per cluster becomes increasingly larger than cost per person we are penalized for adding clusters and the optimal *n* increases. After the optimal *n* is found, the number of clusters can be calculated by plugging back *n* into the formula for *J*:

$$J = \frac{T}{nC_1 + C_2} \qquad \text{[C.4]}$$

The cost per cluster and cost per person may be the same in the control and experimental groups or it may differ. The remainder of this chapter looks at optimal sample allocation when costs of sampling the two groups are equal and when they are not equal.

Equal Costs

The simplest case is when the sampling costs are the same for the treatment and control groups. The following example illustrates how to calculate the optimal *n* and the resulting *J* to minimize the variance for a fixed budget.

A researcher wants to determine the effect of a new drug prevention program in schools. The total budget for sampling costs is \$10000. The cost per cluster (C_2) is \$400 and cost per person (C_1) is \$20. The estimated intra-class correlation coefficient is 0.05. What is the optimal *n*? How many clusters will be in the study? Using formulas 16 and 17 described above, the optimal *n* and *J* can be computed by hand as shown below.

Step 1: Set $\tau + \sigma^2 = 1$, so $\tau = \rho$ and $\sigma^2 = 1 - \rho$. For this example, $\tau = .05$ and $\sigma^2 = .95$

Step 2: Calculate $\sqrt{\tau} = .2236$ and $\sqrt{\sigma^2} = .9747$

Step 3: Find the cost ratio $\frac{C_2}{C_1}$ = 400/20 = 20

Step 4: Set up the equation $n_{opt} = \frac{.9747}{.2236} * \sqrt{\frac{400}{20}} \approx 20$

Plugging 20 into $J = \frac{T}{nC_1 + C_2}$ yields *J* = 12.5 which is rounded down to 12 in order to stay within budget. The value of the variance of the treatment effect can also be calculated by plugging in *n* and *J* to the variance equation.

The Optimal Design software can be used to do these calculations. The software produces a plot as shown below:

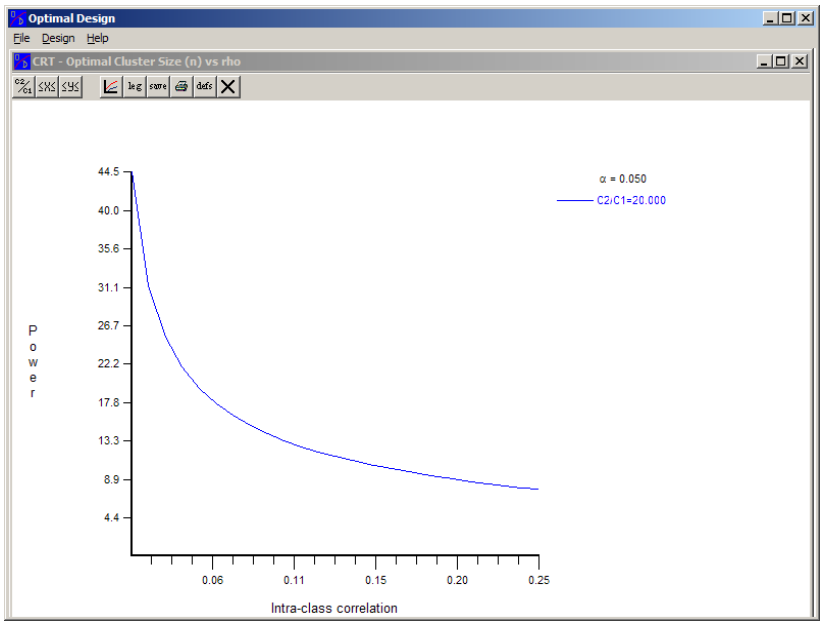


Figure C.1. Optimal allocation curve.

The plot allows the researcher to see how the optimal n changes with respect to the intra-class correlation coefficient. Notice that as ρ increases, optimal n decreases. In other words, if there is large between-cluster variance then it is not very helpful to increase the number of people per cluster and more money should be spent trying to increase the number of clusters.

Notice that in the previous example there were no power calculations or set effect sizes. If the desired effect size is specified, then the Optimal Design software can be used to calculate the optimal *n* and *J* that maximizes power. For example, recall in the example above that: T=\$10,000, C_2=\$400, C_1=\$20, and $\rho = 0.05$. Imagine that the desired effect size is 0.40. Plugging these values into the OD software which solves for *n* and *J* to maximize the power reveals an optimal *n* = 18, *J* =13, and power = 0.53. Knowing that the power is only 0.53 and acceptable power levels are typically 0.80 or higher, the researcher may need to try to increase the budget in order to achieve higher power.

References

Bloom, H.S. (1995). Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs. *Evaluation Review*, 19(5), 547-556.

Bloom, H.S., Bos, J.M., & Lee, S.W. (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. *Evaluation Review*, 23(4), 445-469.

Bloom, H.S., Richburg-Hayes, L., & Black, A.R. (2007). Using Covariates to Improve Precision: Empirical Guidance for Studies that Randomize Schools to Measure the Impacts of Educational Interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30-59.

Bloom, H.S., Hill, C., Black, A.R., & Lipsey, M. (2007). Using Empirical Benchmarks for Interpreting Effect Sizes. *Presentation to the Interagency Roundtable Meeting on "The Application of Effect Sizes in Research on Children and Families: Understanding Impacts on Academic, Emotional, Behavioral, and Economic Outcomes."*

Hedges, L. & Hedberg, E.C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87.

Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, CA: Brooks/Cole Publishing Company, a division of Wadsworth, Inc.

Murray, D.M. & Short, B. (1995). Intra-Class Correlation Among Measures Related to Alcohol Use by Young Adults: Estimates, Correlates, and Applications in Intervention Studies. *Journal of Studies on Alcohol*, 56(6), 681-694.

Raudenbush, S.W. (1997). Statistical Analysis and optimal design for cluster-randomized trials. *Psychological Methods*, *2*, 173-185.
Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W. & Liu, X. (2000). Statistical Power and Optimal Design for Multisite Randomized Trials. *Psychological Methods*, 5(2), 199-213.

Raudenbush, S. W. & Liu, X. (2001). Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Group Differences in Polynomial Change. *Psychological Methods*, 6(4), 387-401.

Schochet, P. (2005). *Statistical Power for Random Assignment Evaluations of Education Programs*. Princeton, NJ: Mathematica Policy Research, Inc.