INSTITUTE OF EDUCATION SCIENCES

July 2014

# Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis


## Authors

Sharon Lohr, Westat
Peter Z. Schochet, Mathematica Policy Research
Elizabeth Sanders, University of Washington



**NCER 2014-2000**
**U.S. Department of Education**

Prepared for:
National Center for
Education Research,
Institute of Education Sciences,
U.S. Department of Education
Washington, DC 20202

Prepared by:
Westat
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Mathematica Policy Research
(Subcontractor)
P.O. Box 2393
Princeton, NJ 08543-2393
(609) 799-3535

# Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis


## Authors

Sharon Lohr, Westat
Peter Z. Schochet, Mathematica Policy Research
Elizabeth Sanders, University of Washington



## National Center for Education Research (NCER)

Meredith Larson (Project Officer)
Phill Gagne

**July, 2014**

Prepared by:
Westat
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Mathematica Policy Research
(Subcontractor)
P.O. Box 2393
Princeton, NJ 08543-2393
(609) 799-3535

NCER-2014-2000

This report was prepared for the National Center for Education Research, Institute of Education Sciences under Contract ED-IES-12-D-0015.

## Disclaimer

The Institute of Education Sciences at the U.S. Department of Education contracted with Westat and Mathematica Policy Research (subcontractor) to develop a paper on partially nested randomized controlled trials in education research. The views expressed in this report are those of the authors and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.



**July 2014**



Lohr, S., Schochet, P.Z., and Sanders, E. (2014). Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis. (NCER 2014-2000) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the Institute website at http://ies.ed.gov/.

**Alternate Formats**

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-0852 or 202-260-0818.

## Disclosure of Potential Conflict of Interest

Westat Inc. is the prime contractor for the NCER Analysis and Research Management Support project, with subcontractors Mathematica Policy Research Inc., and Plus Alpha Research & Consulting, LLC.

**This page left blank for double-sided copying.**

# Acknowledgments

The authors are grateful to Duncan Chaplin, Elaine Carlson, and Jill Feldman for their careful reading of this paper, as well as to the external reviewers for their helpful suggestions. We also thank the following individuals for their contributions: Xiaoshu Zhu provided invaluable assistance for developing the R programs in Appendix C, and Xiaoshu Zhu and Sharon Hirabayashi suggested improvements to increase the efficiency of the SAS code; Tamara Nimkoff managed the project; Lisa Walls, Kevin Collins, William Garrett, and Bonnie Harvey completed the 508 compliance and formatting work; Deirdre Sheehan helped with figures and the cover page; Chantell Atere provided word processing assistance; Steve Bruns and Katie Apolinario helped with project management; and Sharon Clark and Leah Hackleman-Good helped with editing.

**This page left blank for double-sided copying.**

# Table of Contents

**Chapter** **Page**



**Appendixes**

## Appendixes



## Tables

## Tables



## Figures

**Figures** **Page**

**This page left blank for double-sided copying.**

# Introduction 1

Suppose an education researcher wants to test the impact of a high school drop-out prevention intervention in which at-risk students attend classes to receive intensive summer school instruction. The district will allow the researcher to randomly assign students to the treatment classes or to the control group. Half of the students (the treatment group) are assigned to one of four summer classes being offered. The other half (the control group) are not assigned to receive any services during the summer. Thus, the researcher knows there are four clusters in the treatment group: students in the same class share the same teacher and environment and, therefore, are expected to have more similar outcomes than students from different classes. The students in the control group, however, are not assigned to any classes. How are data for the treatment and control group students to be treated in the analysis?

This scenario is an example of a Partially Nested Randomized Controlled Trial (PN-RCT) where treatment students receive intervention services in groups but where this grouping does not occur for control students. The purpose of this paper is to provide guidance to education researchers on how to recognize, design, and analyze data from PN-RCTs to rigorously assess whether an intervention (such as a curriculum, policy, or tutoring program) is effective.

Chapters 1 and 2 of the paper are written primarily for applied education researchers with an introductory knowledge of quantitative impact evaluation methods. Our goal is to help these researchers negotiate key concerns when proposing and conducting research using PN-RCT designs. The paper addresses design issues such as possibilities for random assignment, cluster formation, statistical power, and confounding factors that may mask the contribution of the intervention. Chapter 3 is intended for education researchers interested in estimating treatment effects for PN-RCT designs; it discusses basic statistical models that adjust for the clustering of treatment students within intervention clusters, associated computer code for estimation, and a step-by-step guide, using examples, on how to estimate the models and interpret the output. Chapter 4 and the technical appendixes discuss more advanced statistical topics pertaining to PN-RCTs and are written primarily for an audience with a strong statistical background.

In the remainder of this introductory chapter, we define PN-RCTs, with specific examples that are described in the broader context of the choices for research designs, and provide a roadmap to the rest of the paper and a summary of our take away messages.

## 1.1 What Is a PN-RCT?

To clarify the concept of a PN-RCT, let's consider the distinction between clustered (nested) and unclustered (non-nested) randomized trials. In some educational interventions, individual students are randomized directly to the treatment or control group, and both intervention and control protocols are administered in an individual setting. Such an experiment is an Individual-Level Randomized Controlled Trial (I-RCT). An example of an I-RCT would be an experiment with home-schooled students in which students in the treatment group are given a tablet computer with an adaptive learning program and where students in the control group use the standard curriculum without the extra tablet computer. The random assignment to research groups is done at an individual level. If 120 students participate in the study, the randomization can be performed by placing 60 red balls and 60 white balls in a box and drawing balls without replacement to assign students in the list of eligible participants to the control (white ball) or treatment (red ball) group. There is no clustering in either the control or treatment arm of the study. Figure 1 depicts an I-RCT.

In many educational studies, interventions are instead administered at the classroom, school, or district level. Consider an experiment to evaluate the effects of a curriculum in which music is used to help students understand fractions. Suppose that 20 math classes, each with 25 students, are available to participate in the study. In a Cluster Randomized Controlled Trial (C-RCT), 10 classes are randomly assigned to the treatment group (where music is used), and the other 10 classes are assigned to the control group (where music is not used). In this design, students are *nested* in classes: each student belongs to exactly one class, and students in the same class *all* receive the same music instruction (either treatment or control). Thus, although 500 students participate in the study, there are only 20 classes, and, because of the way the randomization was done, classes are the units of analysis. In a C-RCT, the pre-intervention characteristics of study teachers and their students will be balanced on average across all possible values for the treatment and control classrooms, although they could differ numerically in any given RCT due to random sampling. The outcomes of students in the same class are expected to be positively correlated because they share the same teacher and

environment, and this correlation must be accounted for in the data analysis to arrive at correct conclusions about the effect of the intervention. Figure 2 depicts a C-RCT. [1]

**Figure 1. Depiction of an I-RCT**

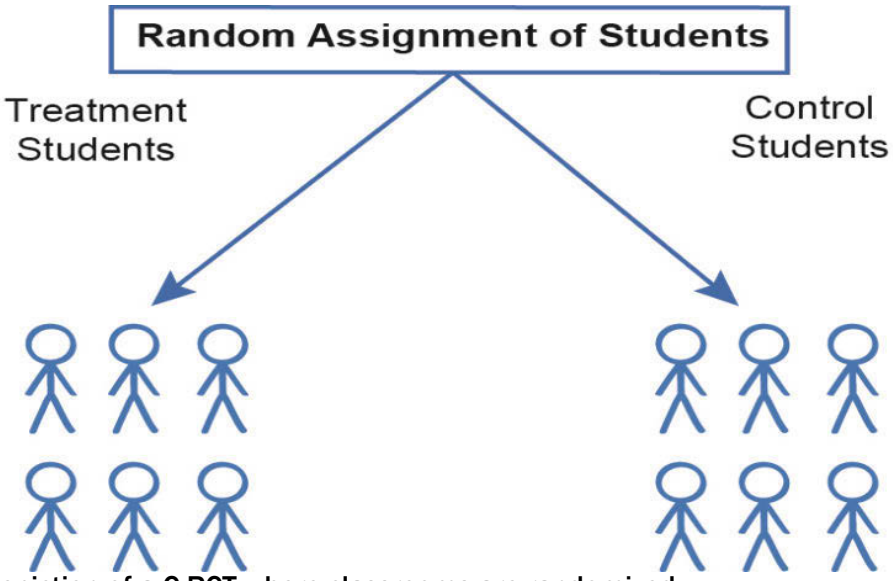
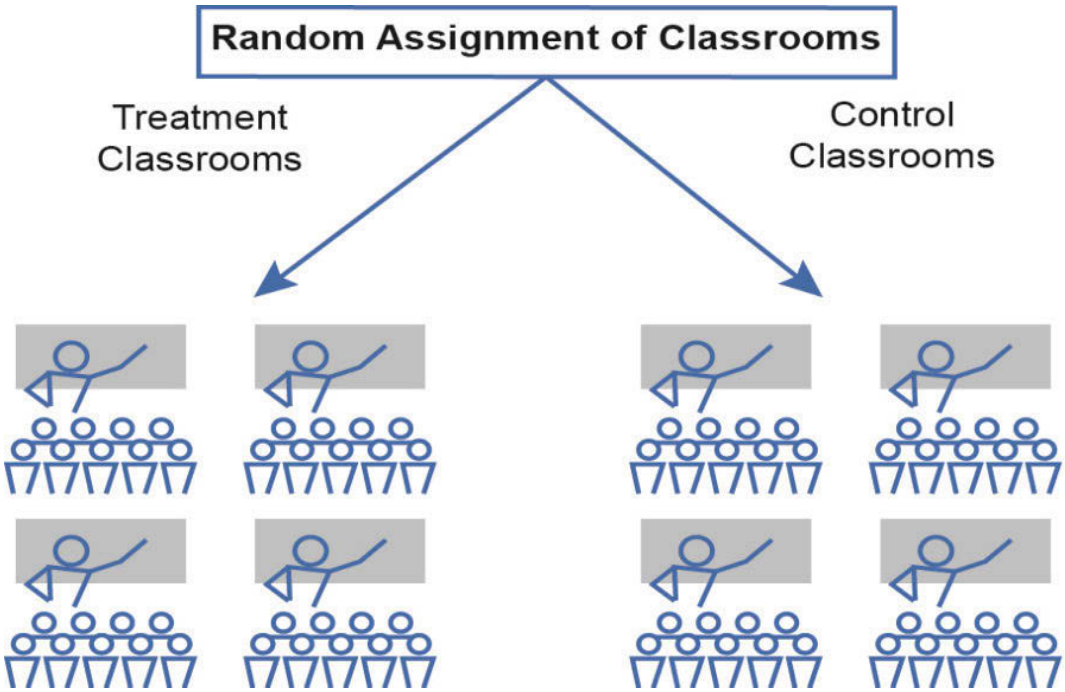
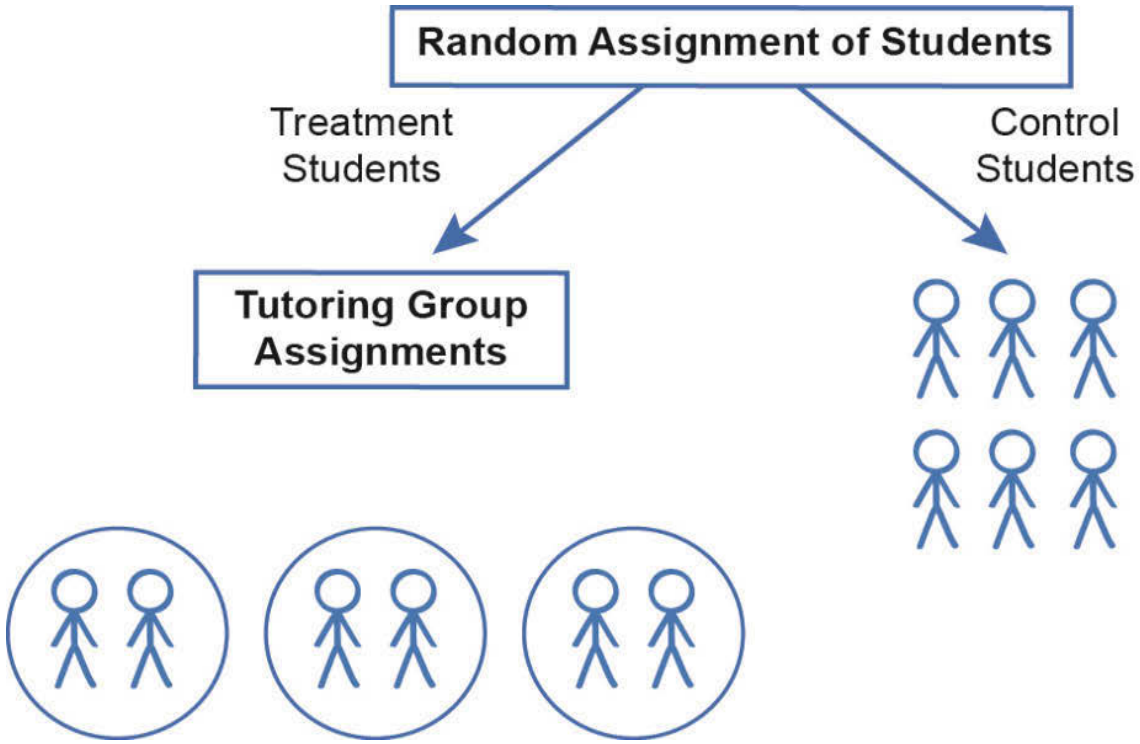



**Figure 2. Depiction of a C-RCT where classrooms are randomized**




[1] In this C-RCT example, random assignment could be conducted using intact classrooms. Alternatively, students could be randomly allocated to classrooms prior to the random assignment of classrooms to increase the precision of the estimated treatment effects. Both of these designs are C-RCTs due to cluster-level random assignment.

Figure 3. Depiction of a PN-RCT for a small group tutoring intervention




A PN-RCT is a hybrid of an I-RCT and a C-RCT. In a PN-RCT, students in the treatment group are clustered like those in a C-RCT, and students in the control group are unclustered like those in an I-RCT. The design is called *partially nested* because students in the treatment group are nested in some higher level unit, such as a tutoring group or class, but students in the control group are not nested as part of the experimental design. Because the design combines features of both an I-RCT and a C-RCT, the analysis must combine features of both as well. Figure 3 depicts a PN-RCT.

PN-RCT designs are common in educational and behavioral research where interventions are often delivered in group settings, whereas the control protocol often involves no additional clustering. In a literature review of public health research, Bauer, Sterba, and Halfors (2008) found that 32 percent of the 94 randomized experiments identified across four journals from 2003-2005 used PN-RCT designs. In a similar literature review of educational research, Sanders (2011) found that about 15 percent of the 75 randomized experiments identified across four journals from 2007-2009 used PN-RCT designs. [2]

[2] These journals included the *American Educational Research Journal*, the *Journal of Educational Psychology*, *Contemporary Educational Psychology*, and *Remedial and Special Education*.

Although much has been written about how to address design and analysis concerns in I-RCTs and C-RCTs (see, for example, Donner and Klar 2000, 2004; Raudenbush and Bryk 2002), there is less guidance on how to navigate these issues for PN-RCTs. This paper provides such guidance.

## 1.2 Recognizing Education PN-RCTs

PN-RCTs are experimental designs where treatment group students receive intervention services in clusters—referred to as *intervention clusters* (ICs)—but where this type of clustering does not affect those in the control group. In a PN-RCT, individuals in the same IC may share common characteristics and intervention experiences because they are exposed to the same tutor, teacher, curriculum, or learning environment.

### 1.2.1 The Basic PN-RCT Design

The simplest PN-RCT design is where individuals from a *single population*—such as a single school, school district, Head Start center, or summer program location—are randomly assigned to experimental conditions. It is assumed that treatment group individuals are subsequently placed into ICs, whereas control individuals are not. For this design, the only source of clustering is due to the ICs created by the intervention; additional sources of clustering due to other nesting structures are not present.

Sometimes it is difficult to recognize a real-life experiment as having the basic PN-RCT structure because real experiments often blend features from several types of designs. We start, then, with the hypothetical experiment briefly mentioned in section 1.1 on using music to help students understand fractions and proportional reasoning and discuss how it might be alternatively conducted as a C-RCT, as an I-RCT, and as a PN-RCT.

**Example 1.1. Three Alternative Designs for an Experiment.** The experiment on using music described in the introduction is a classic C-RCT. The students are clustered in 20 classes; half of the classes are randomly assigned to the treatment group, and the other half are assigned to the control group. The teachers in the treatment group classes have the students study songs that illustrate different concepts about fractions: the Blue Danube Waltz for thirds and Twinkle Twinkle Little Star for halves and fourths, and working up to Mars: Bringer of War (from The Planets by Gustav

Holst) for fifths and The End (from Abbey Road by the Beatles) for combining fractions and switching from thirds to fourths.

In the C-RCT, the students in both the control and treatment groups are clustered. There are many potential reasons that students in the same class will tend to perform similarly on the assessment: students in the same class may have similar backgrounds, they are all being taught by the same teacher (and teachers may vary in effectiveness), and the group setting means that some students can influence the performance of their peers. This experiment, therefore, needs to be analyzed in the same way as it was randomized. The 20 classes can be considered independent of each other because they are the units being randomized, but individual students are positively correlated with other students in the same class because of the shared environment and teacher. Thus, when analyzing the data, there are only 20 independent units (the classes). The data can be analyzed using hierarchical linear modeling (HLM) methods at the student level. Alternatively, a simple but correct analysis can be performed by calculating the mean test score for each class so that there are 10 observations for the control group and 10 observations for the treatment group and then using a two-sample t test with the mean scores as the 20 observations for the experiment.

It is more difficult to visualize conducting this study as an I-RCT. To do that, the intervention would need to be in a form that is administered in an individual setting rather than in a group setting. Let's suppose, then, that the music curriculum is placed on a tablet computer. There is a pool of 500 students available for the study. Half of them are randomly selected to get the music curriculum to work on by themselves at home, and the other half in the control group do not have any additional activities. The students are not clustered in any way in terms of receiving the intervention or control protocols; students in both groups can be considered to be independent observations. The data can be analyzed by using a two-sample t test with 250 students in each group.

For a PN-RCT, as in the I-RCT, there is a pool of 500 students available, and half are randomly assigned to receive the treatment protocol and the other half to the control protocol. But in the PN-RCT design, the students in the treatment arm receive the intervention in a group setting: they are grouped into 10 ICs of 25 students each, with each IC taught by a different teacher. The students in the control arm have no such clustering. The analysis of the experiment needs to treat the students in the control arm as independent observations but needs to account for the clustering in the treatment arm. Methods for doing this are described in chapter 3.

The key features of a basic PN-RCT design are

- Individual students are randomly assigned to the treatment or control arm of the experiment.
- The intervention is administered in a group setting, and the ICs are formed after the students have been assigned to the treatment arm. ICs can be formed purposively by educators (e.g., based on characteristics of students and teachers) or could be formed randomly.
- At least two ICs must be formed, so it is possible to calculate variances of the estimated intervention effects.
- The control protocol is administered to individuals who are not in ICs.

Thus, the basic PN-RCT design is typically used for interventions that take place outside of the typical classroom setting: summer programs, experiments for improving home-schooling, pre-school interventions, or after-school programs. The design may also be used for a study within a single school, provided that students within the school are randomized *individually* to the treatment and control groups and that the ICs are formed after randomization (that is, the ICs are not the same as the regular classrooms). Example 1.2 describes a real-world education study that fits the structure of a basic PN-RCT.

**Example 1.2. Pre-K Social-Communication Intervention for Children with Autism.** Roberts et al. (2011) investigated the effects of an early social-communication intervention program for preschool children diagnosed with an autism spectrum disorder. Children were recruited from the local area and were randomly assigned to one of two conditions: (1) a small-group, center-based intervention or (2) an individualized, home-based intervention. The center-based (treatment) condition (n = 29) involved groups of five to six children (with a simultaneous parent support group), whereas the home-based condition (n = 27) took place in the child's home with the parent present. Each child and parent was assessed before and after the intervention was delivered over a 40-week study period. This study has the key features of a basic PN-RCT: (1) individual children were randomly assigned to the treatment and home-based groups, (2) children randomized to the treatment group were then placed in centers (ICs) and received the intervention in a small group setting, and (3) children randomized to the home-based group received the home-based intervention individually (not in a group setting).

The next hypothetical example describes a design that appears at first glance to be a PN-RCT but is in fact not a valid study design. This example explains why at least two ICs are needed for the design.

**Example 1.3. An Invalid Study Design.** Consider a hypothetical experiment conducted at one site in which half of the students are randomly assigned to the treatment group and the other half are randomly assigned to the control group. The students randomly assigned to the control group receive no intervention; the students in the treatment group all attend the same tutoring session (IC). Note the difference from the PN-RCT design in Example 1.1: in Example 1.1, 10 ICs were formed, while in this example, only 1 IC is formed. Here, the effect of the intervention cannot be separated from the effect of the treatment students' classroom or teacher. The students in the same IC may be positively correlated, but that correlation cannot be estimated in this experiment. For variance estimation purposes, the sample size in the control arm is the number of students in the control group, but the sample size in the treatment arm is one: the number of ICs. Stated differently, with only one IC, it is not possible to estimate a variance for the treatment group because this variance represents the extent to which *mean* student outcomes vary across the ICs. Thus, with only one IC, it is not possible to conduct a two-sample *t* test.

This same problem occurs in many fully nested experiments in education, particularly in small-scale studies. A researcher has two classrooms available, and randomly assigns one classroom to the control group and the other to the treatment group. Essentially, the sample size is one (the number of classes) in each treatment arm, and a sample size of one does not allow you to estimate variability. This is an improper study design and does not allow for inferences to be made about the effectiveness of the intervention. To make valid statistical inferences from a fully nested design, you need to have at least two clusters in each study arm.

The same principle carries over to PN-RCTs. At least two ICs are needed to make inferences about the effectiveness of the intervention. The multiple ICs can be in the same site, as described in Examples 1.1 and 1.2. Alternatively, multiple sites or blocks may be used, as described in section 1.2.2.

### 1.2.2 The Blocked PN-RCT Design

It is often the case that a single site will not have enough students to allow the effect of the intervention to be detected with sufficient power. In addition, experiments done at a single site may have limited generalizability to other sites. In a *blocked*[3] *PN-RCT* design, the basic PN-RCT design is

[3] Some researchers refer to this type of design as a *stratified* design and use the term *strata* instead of *blocks*. In the following, we use the term *block* to refer to a site in which the basic PN-RCT design is replicated.

replicated separately across several sites (blocks), such as cities, schools, or classrooms. In this design, students are randomized to treatment and control groups separately in each block.

The key features of a blocked PN-RCT design are

- Potential participants in the experiment are arranged in blocks before randomization takes place. These blocks can be naturally occurring units such as classrooms, schools, districts, or cities. Alternatively, blocks may be deliberately formed by the researcher in advance of the study. For example, the researcher may group students into three blocks by their score (high, medium, low) on a pretest.

- Individual students are randomly assigned to the treatment or control arm of the experiment. The randomization is carried out *separately* within each block. Thus, if blocks are schools, half of the participating students within School A are randomly assigned to the treatment group, and the other half are assigned to the control group. The same randomization procedure is carried out for School B, for School C, etc.

- The ICs are formed for the treatment students separately within each block. Thus, if blocks are schools, the students randomized to the treatment group from a school are then assigned to one or more ICs; each of those ICs contains students from only that school (i.e., ICs are nested in schools). More complex designs, where ICs contain students from more than one block, are described in section 4.2.

- In contrast to the single-site PN-RCTs described in Examples 1.1 and 1.2, in the blocked PN-RCT it is permissible to have one IC in each block, as described in Example 1.5; the multiple blocks provide the replication needed to make valid conclusions from the study.

- The control protocol is administered to individuals randomized to the control group in each block.

Because the randomization is done separately in each block, the comparison group for the treatment students in a block is the set of control students in that block.

**Example 1.4. Elementary Summer Program for Disadvantaged Children.** Chaplin and Capizzano (2006) conducted a random assignment evaluation of the Building Educated Leaders for Life (BELL) program—a summer school program designed to improve academic skills and social behaviors among low-income, academically challenged children. In this study, more than 1,000 elementary school children who applied to the program in New York and Boston in 2005 were randomly assigned to either a treatment group that was selected to participate in the program or a control group that was not. For academic activities, students selected to participate in the program were grouped into more than 30 ICs of about 15 children each, where each IC was taught by a regular teacher from the public schools and an experienced teaching assistant.

This design classifies as a blocked PN-RCT design because (1) students were randomly assigned to a treatment or control condition within each city (blocking units), (2) treatment students were grouped into summer school classes using BELL placement rules in operation at the time of the study, and (3) there is no comparable clustering for control students who were not offered summer school slots.

**Example 1.5. An After-School Intervention.** A PN-RCT design similar to the BELL evaluation was used for the impact evaluation of the 21st Century Community Learning Centers program (James-Burdumy et al. 2005). This evaluation was conducted in 26 after-school centers in 12 school districts, where the after-school programs typically offered homework sessions, academic activities, enrichment activities (such as art, drama, or music), and recreation activities. For the elementary school design, more than 2,300 elementary school students interested in attending a *specific* 21st Century Center were randomly assigned to a treatment group that could attend the center or a control group that could not.

Random assignment was conducted within each center, so this design is a blocked PN-RCT. The blocks consist of the students interested in attending each center. Within each block, the students randomly selected to attend the center form the IC, and the students not selected to attend the center form the control group.

Note that if the students attending the center are considered to be the IC, there is one IC per block in this design. Because there are 26 blocks, however, the treatment effect can still be estimated from this design (unlike the setting in example 1.3 where there is only one IC in the experiment).[4]

Examples 1.4 and 1.5 describe blocked PN-RCTs in which the intervention takes place outside of the regular school setting. In Example 1.6, students in the treatment group are pulled out of school-day activities to participate in the intervention.

**Example 1.6. Kindergarten Math Intervention.** A smaller-scale example of a blocked PN-RCT design was used by Dyson, Jordan, and Glutting (2013) to investigate an intervention intended to help kindergarteners develop core number competencies. Participants were recruited from five schools with large numbers of students at risk for mathematics difficulties. Children were

[4] Consider the analogy with a simpler experiment. If an experiment has only two students—one in the control group and one in the treatment group—then the effect of the intervention cannot be evaluated because there is no information on the variability in each group. If, however, 10 students are formed into 5 pairs, and 1 student in each pair is randomly assigned to the control group and the other to the treatment group, then a paired *t* test can be conducted to evaluate the effects of the intervention.

randomized separately within each kindergarten classroom to the treatment group or to the control group. Both groups of children received the usual mathematics instruction during their regular mathematics instruction period. Children in the treatment group were placed into groups of 4 for 30 minutes per day, 3 days per week for 8 weeks; the intervention took place at a time when the children were not receiving their regular math instruction and was conducted at a small table either inside the classroom or just outside the classroom.

This example is a blocked PN-RCT with classrooms serving as the blocking unit. The design classifies as a PN-RCT because (1) children are randomly assigned to the treatment or control group within each classroom, (2) the control group students receive no additional instruction, and (3) the treatment group students receive the intervention in a small group setting.

### 1.2.3 C-RCTs That Share Features of a PN-RCT

Other forms of designs also fall under the PN-RCT umbrella. These include designs in which more complicated forms of nesting occur along with the basic PN-RCT asymmetry between the treatment and control groups.

As an example, let's consider a design where *schools* are randomly assigned to treatment and control groups, and then ICs are formed in the treatment schools. Is this a PN-RCT? Such an experiment is technically a C-RCT because schools rather than individual students are randomly assigned to the treatment and control groups. Thus, school-level clustering applies to *both* the treatment and control groups. But the treatment group has an additional source of variability (the ICs) not found in the control group.

These types of "hybrid" C-RCT designs share features of PN-RCTs in the sense that clustering effects *differ* for the treatment and control groups due to the ICs. Thus, we consider these designs as falling into the general class of PN-RCTs because a different design structure exists for the two research groups. We consider statistical aspects of these designs in chapter 4, which, for simplicity, we label as "PN-RCTs" even though this is a bit of a misnomer because these designs have clustering in both research arms but with an additional source of clustering in the treatment group due to the ICs.[5] Example 1.7 illustrates such a design in which students are pulled out of regular classrooms in the treatment schools but not in the control schools.

[5] As discussed in Chapters 3 and 4, the intraclass correlation coefficients (ICCs) that are often used to examine clustering effects in C-RCT designs become more complex for hybrid C-RCT designs because the ICCs for the treatment group must reflect multiple layers of clustering. Thus, in these designs, the ICCs will differ for the treatment and control groups.

**Example 1.7. Pull-Out Mathematics Program.** In the Evaluation of the Number Rockets Intervention (Rolfhus et al. 2012), 76 schools in 4 urban districts were randomly assigned to a treatment or control condition. In the treatment schools, grade 1 students at risk for difficulties in mathematics were provided intensive mathematics instruction by a tutor—in small groups—that met outside the classroom during the regular school day (but not during regular mathematics classes). These pull-out groups formed the ICs.

This design is a C-RCT because schools were randomized. However, this design can also be considered to be in the class of PN-RCT designs because students in the treatment schools were placed into small tutoring groups (the ICs). Thus, clustering effects differed for the treatment and control groups: for the treatment group, students were nested within ICs that were nested within schools, whereas for the control group, students were nested only within schools. This asymmetric design structure should be taken into account in the analysis (see section 4.1).

We have discussed the use of schools both as potential blocking units (section 1.2.2) and as potential clusters (this section). In both situations, students who are in the same school are expected to be more similar than students who are in different schools because they share the same school environment, may live in the same neighborhood, or may have other factors in common. A researcher might ask, then, why a school is considered a blocking unit for some experiments and a clustering unit for others. The answer depends on *how the randomization to the treatment and control groups is performed*. If students are randomly assigned to treatment and control groups separately *within* schools so that each school contains both treatment and control students, then schools are a blocking unit. In that case, the similarity of students in the school helps increase the precision of the estimated treatment effect because the treatment students in the school are compared with the control students in that school, who share the same environmental factors. If entire schools are randomly assigned, however, with some schools receiving the treatment protocol for all of their participating students and the others receiving the control protocol for all of their participating students, then schools are clusters. The treatment students are in completely different schools from the control students, so the estimated treatment impact includes the school-to-school variability as well as the student-to-student variability. In general, blocking increases precision, while clustering decreases it.

## 1.3 Contrasting I-RCTs, C-RCTs, and PN-RCTs

Much has been written about the design and analysis of data from C-RCTs and the distinction between an I-RCT and a C-RCT. Useful references for C-RCTs include Donner and Klar (2000, 2004), Campbell, Elbourne, and Miller (2004), and Raudenbush and Bryk (2002). In this section, we summarize similarities and differences among I-RCT, C-RCT, and PN-RCT designs and analyses. Table 1 outlines the major distinguishing features of the three designs.

RCTs are considered the gold standard[6] for education evaluations that address causal questions because, in expectation, the randomization process divides measured and unmeasured factors that could influence the outcome approximately evenly between the treatment and control groups. Thus, in an I-RCT, we would expect to have about half of the highly motivated students in each study arm; in a C-RCT, the classrooms or schools with superior teachers are equally likely to be in the treatment or control group, and the academic abilities of students in the randomized clusters will be balanced across the two research groups. Ideally, in an RCT, the only difference between the control group and the treatment group is the presence of the control or intervention protocol. This feature of a randomized study allows us to conclude that an intervention *caused* a change in student outcomes.

A PN-RCT is a special form of an RCT in which the intervention is administered in a group setting but the control protocol is not administered in a group setting. The basic feature of an RCT—randomization of units to the treatment or control condition—still holds. Thus, we would expect that the randomization process would divide other characteristics approximately evenly between the treatment and control groups so that the groups differ only by the presence of the intervention, which is administered in a PN-RCT to clusters of students. Importantly, PN-RCTs are RCTs whether or not treatment students are randomly assigned to ICs. IC formation has no effect on the overall comparability of the treatment and control groups, which is the defining feature of an RCT, although there are advantages and disadvantages to whether ICs are formed purposively or randomly (as discussed in more detail in chapter 2).

[6] See *ies.ed.gov/ncee/pubs/evidence_based/**randomized**.asp*

Table 1. Contrasting features of I-RCT, C-RCT, and Basic PN-RCT designs

| | I-RCT | C-RCT | PN-RCT |
|---|---|---|---|
| Randomization | Individual students are randomized to treatment and control groups | Clusters of students are randomized to treatment and control groups. | Individual students are randomized to treatment and control groups. Students in the treatment group are further assigned in some way to the different ICs. |
| Cluster formation | No clusters | Usually naturally occurring clusters such as classrooms, schools, or school districts | ICs are formed in the treatment group only after students are assigned to study arms. |
| Independent units | Students | Clusters | Students in the control arm; ICs in the treatment arm |
| Typical statistical method used to evaluate intervention impact using a continuous outcome measure | Two-sample *t* test (or nonparametric test such as Wilcoxon) | Hierarchical linear model, with clusters as second-level units and students as first-level units | Special form of HLM, described in chapter 3 |

## 1.4 Roadmap to the Rest of the Paper and Take Away Messages

The remainder of this paper addresses statistical design and analysis issues for PN-RCTs. Chapter 2 reviews some of the statistical concerns applied researchers may have when preparing to conduct a PN-RCT. Chapter 3 contains more detailed information about how to analyze data collected via the simplest type of PN-RCT design and the blocked PN-RCT design. This chapter presents the statistical models for the designs and provides SAS code, output, and interpretation of the results for data examples.

Chapter 4 discusses advanced topics and is written primarily for an audience with a strong background in HLM methods. In some experiments, PN-RCT features may be incorporated as part of a larger study that involves multiple layers of clustering for both the treatment and control groups. Chapter 4 discusses these alternative designs and shows how to analyze data from them using SAS software. It also discusses statistical issues of power calculations for the basic and blocked PN-RCT designs discussed in chapters 1-3 as well as the designs discussed in section 4.1.
The appendices discuss other topics that may be of interest to researchers. Appendix A, intended for readers interested in the statistical formulation of the models, places PN-RCTs in the general

context of a mixed linear model. Appendix B discusses issues of degrees of freedom calculations. Appendices C and D show how to analyze data from a simple PN-RCT using R and HLM software, respectively. Finally, Appendix E gives the complete SAS code for generating the data sets used in the examples and analyzing the results.

We focus on the design and analysis of PN-RCTs for a continuous post-intervention measure, such as an achievement test score, that is analyzed at the student level. Our results also apply to binary outcomes (for example, high school graduation status or proficiency in math or English) that are analyzed using linear models. We do not discuss nonlinear or generalized linear models (such as probit or logit models) for analyzing binary or ordinal outcomes. However, the basic concepts presented in this report also pertain to nonlinear models.

The report focuses on statistical aspects of PN-RCTs. We do not discuss other important evaluation design issues, such as informed consent and other ethical issues for experimental designs and how they might differ for PN-RCTs. For example, PN-RCTs might involve collecting data on teacher-student links that would otherwise be confidential and not required in other types of RCTs. In this case, if students are tracked into high and low-ability groups, parents might not want this information to be revealed. The National Forum on Education Statistics (2010) Guide to Data Ethics discusses ethical issues and guidelines for education research.

To help the reader sift through the detailed analyses presented in this paper, we first summarize our key points, by topic, as follows:

**Defining PN-RCTs (chapters 1 and 2)**

**PN-RCTs occur when the ICs are groupings formed specifically for the study, apart from naturally occurring clusters such as classes or schools.** In an education PN-RCT, the intervention must supplement the status quo learning environment and be administered in a group setting. The formation of ICs for the treatment group leads to an asymmetric design structure for the treatment and control groups; this asymmetry is the *defining* feature of PN-RCTs. For simplicity, we label designs as "PN-RCTs" even if more complicated forms of nesting occur along with the basic asymmetry between the treatment and control groups due to the ICs.

**PN-RCTs are randomized experiments regardless of how ICs are formed.** Because of random assignment, in expectation, the *full* treatment and control groups will be balanced in terms of their baseline characteristics (both observed and unobserved). Consequently, the comparison of the mean

outcomes across the two groups provides unbiased estimates of the causal effects of the intervention on key student outcomes. This property holds regardless of how treatment students are assigned to ICs.

**Broad Design Considerations for PN-RCTs (chapter 2)**

**In PN-RCTs, ICs present in the treatment condition but not the control condition may introduce confounds that could bias the impact findings.** Care must be taken when designing PN-RCTs to ensure that the differences between the treatment and control groups are due to the intervention and not some other extraneous factors that correlate with study outcomes and treatment status. Confounds may exist because the treatment students receive intervention services in a new group environment (the ICs), whereas the control students do not. For instance, the ICs may make it easier to administer the evaluation for the treatment group (e.g., obtaining signed study consent forms and collecting data) than for control group students who may be more dispersed and less invested in the study. As another example, grouping students into small ICs may be an intervention in itself.

**The way in which students are assigned to ICs should conform to the study research questions and implementation context.** In some PN-RCTs, researchers may not have the flexibility to design the way ICs are formed; this might occur, for example, in evaluations of existing or ongoing interventions where educators have well-established mechanisms for forming the ICs. In other PN-RCTs, researchers may not want to alter the IC formation process (even if they could) so that they can examine the effects of interventions as typically implemented. In some evaluations, researchers may want to randomly assign students to ICs to be able to rigorously compare teacher practices across ICs.

**PN-RCTs can accommodate multiple treatment groups.** In some PN-RCTs, education researchers might want to test the relative effects of several intervention components in isolation or in combination, so study results can be used to develop a promising package of intervention services. For example, an education researcher may be interested in varying intervention dosage levels across ICs or in testing different intervention features (e.g., IC size and curriculum) in a factorial design. These multi-armed PN-RCTs have the same general design structure as the PN-RCT with a single treatment and control group because they share the common feature that treatment students are nested within ICs.

**Many PN-RCT issues apply also to quasi-experimental designs (QEDs).** A partially nested design structure can also exist in QEDs, such as matched comparison group, regression discontinuity (RD), instrumental variable (IV), or single-subject designs. These designs all share the common feature of a treatment group that receives intervention services and some comparison group that does not. Thus, a partially nested design structure emerges if the treatment group receives intervention services in clusters, whereas IC-level clustering does not exist in the control group.

**Statistical Power Considerations for PN-RCTs (chapters 2 and 4)**

**Sample sizes must be somewhat larger under PN-RCT designs than under traditional RCT designs without ICs to achieve precise impact estimates.** IC effects typically increase the variances of the responses for the treatment group, so required student, classroom, and school samples must be larger in PN-RCTs than in traditional RCTs without ICs to achieve the same level of statistical precision. For similar reasons, specialized sample size formulas are required for PN-RCTs.

**Precision levels can typically be improved if more ICs and fewer students per IC are sampled for the study.** Subject to study resource and implementation constraints, researchers designing PN-RCTs should be aware that statistical power can be improved by selecting more ICs with fewer students per IC. This design could be implemented, for example, by randomly subsampling students within ICs to allow for a larger sample of ICs.

**An important area for future research is to obtain empirical values for intraclass correlation coefficients (ICCs) for PN-RCT designs.** This report presents sample size calculations assuming a range of plausible values for clustering effects due to IC formation. An important area for future research is to use a range of datasets and alternative outcome measures to obtain empirical values for these ICCs in multiple settings, so education researchers can use appropriate values when planning their PN-RCTs.

**Statistical Analysis for PN-RCTs (chapters 2, 3, and 4)**

**The asymmetric design structure in PN-RCTs complicates the estimation of treatment effects and the computer code needed for estimation.** In PN-RCTs, the variances of the treatment and control group mean outcomes could differ. This asymmetry occurs because IC effects pertain to the treatment group but not to the control group. This leads to a different model error

structure for the treatment and control groups. SAS, R, and HLM code needs to be adapted to meet the programming requirements of PN-RCTs.

**In this paper, we focus on methods for analyzing data from PN-RCTs that treat ICs as random factors in the impact estimation models.** There is some controversy in the literature about whether one should regard variability across ICs as a *random* factor for purposes of generalizing the impact findings to broader IC realizations or whether one should regard ICs as *fixed*, thereby restricting inferences to the particular ICs used in the study. In reality, some IC-related factors are likely to be random, and others are likely to be fixed; hence, treating all IC-related factors as fixed could result in underestimation of the variance of the treatment impact. Thus, in this paper, we take a more conservative approach and provide methods for analyzing data from PN-RCTs that treat ICs as random factors. The approach one adopts, however, is based on assumptions.

**The inclusion of detailed baseline covariates in the impact estimation models can help unify various statistical approaches for treating IC effects in the analysis.** If students are randomly assigned to ICs, the various approaches for analyzing PN-RCT data are similar because variability across ICs will be minimized. However, if students are tracked into ICs based on their pre-intervention characteristics, it is critical that baseline covariates be used in the models to help explain this tracking. For PN-RCTs, researchers must recognize the importance of collecting information on IC assignment rules and collecting baseline data to measure these rules.

# Key Design and Analysis Issues for PN-RCTs 2

This chapter reviews some concerns that applied education researchers may have when preparing to conduct a PN-RCT in terms of appropriate design and analysis. We discuss topics such as IC formation, sample size requirements, data needs, and background issues on statistical estimation. The presentation is intended to be conceptual and nontechnical and to build off the presentation in chapter 1. A more detailed statistical treatment of estimation methods for PN-RCTs is provided in chapters 3 and 4.

A rigorous PN-RCT evaluation must adhere to basic principles of good scientific practice. First, the evaluation research questions should drive the design decisions and not the other way around. Study research questions should be based on a clear conceptual model of intervention components and the hypothesized causal chain leading to expected intervention effects on key mediating and longer term outcomes. As we shall see, PN-RCTs can be designed to address different research questions.

Second, a PN-RCT needs to be designed with the analysis in mind. The basic rule in analyzing randomized experiments is: *Analyze as the study was designed.* In a PN-RCT, the analysis methods must account for the different design structure for the treatment and control groups due to IC formation. This leads to statistical models that are more complex than for standard RCT designs where the design structure is symmetric for the treatment and control groups.

Importantly, PN-RCTs are randomized experiments regardless of how ICs are formed. In PN-RCTs, students are *first* randomized to treatment and control groups, and *second*, treatment group students are allocated to ICs using some placement mechanism. Thus, in expectation, the *full* treatment and control groups will be balanced in terms of their baseline characteristics (both observed and unobserved). Consequently, the comparison of the mean outcomes across the two groups provides unbiased estimates of the causal effects of the intervention on key student outcomes.

## 2.1 Design Questions for PN-RCTs

PN-RCTs should in general be considered only when the ICs are groupings formed specifically for the study, apart from naturally occurring clusters such as classes or schools. If the intervention is to be administered to entire classes that already exist, then the appropriate control group consists of other classes that already exist, and a C-RCT should be used. As discussed in chapter 1, other forms of designs also fall under the PN-RCT umbrella. These include designs in which more complicated forms of nesting occur along with the basic PN-RCT asymmetry between the treatment and control groups.

PN-RCT designs could be appropriate for testing either *existing* or *new* interventions that involve group administration outside of the normal classroom environment. The feasibility of PN-RCT designs will clearly depend on logistical considerations, such as whether the study schools have sufficient sample sizes to generate precise impact estimates and sufficient space and staff for delivering the group instruction.

To help applied education researchers assess whether a PN-RCT is appropriate for their evaluations, we consider several design-related questions that might help inform this choice:

**Is there a minimum number of ICs that are required for a PN-RCT?** A PN-RCT where IC factors are treated as random must contain at least two ICs to produce proper variance estimates. The appropriate number of ICs for an evaluation, however, will typically be much larger, so the study can produce precise estimates of intervention effects (see section 2.4).

**Is there a minimum number of treatment students per IC that are required for a PN-RCT?** In a "pure" PN-RCT where IC factors are treated as random, each IC will contain at least two treatment students. There may be evaluations, however, where ICs have only one student. This could occur, for example, because some students with special needs require their own tutor or because some students have missing outcome data. If *every* IC has only one student, the design is an I-RCT, not a PN-RCT. However, designs where some ICs have one student and others have more students fit the PN-RCT structure for part of the treatment group but not for those in the "singleton" ICs. The estimation methods discussed in chapter 3 for the basic PN-RCT design can be easily adapted to account for these singletons.[7]

[7] Specifically, the random IC-level error terms would not be included in the estimation models for these singletons because these observations are independent of other observations.

**What if the intervention is a pull-out program that removes students from their regular classrooms? Is this a PN-RCT?** The answer is "it depends." To help explain this complex issue, it is important to remember that we have defined PN-RCTs to include designs where *other sources of nesting could exist* in both the treatment and control groups but where nesting due to ICs pertains only to the treatment group. Keeping this in mind, consider the following designs:

- Suppose students are randomized within classrooms and that the pull-out program *supplements* the normal classroom instruction. The pool of randomized students can consist of all students in the classroom or targeted students only (for example, English language learners). We assume that the pull-out program meets outside normal classroom hours (for example, after school when the control students are no longer in school) so that the treatment and control group students spend the same amount of time in their regular classrooms. An intuitive way to view this design is that a "mini-experiment" is being conducted in each classroom, where each classroom is its own block (site). Treatment and control students are nested within the same regular classrooms, but the treatment students have an *extra* source of clustering due to the ICs. Thus, this design is a PN-RCT due to this extra layer of clustering for the treatment students. This design should be treated as a blocked PN-RCT where classrooms are the blocks. This design addresses the following research question: Does a supplementary pull-out program improve student outcomes?

- Suppose instead that the pull-out program *fully replaces* the normal classroom instruction. In this case, treatment and control students would attend different classrooms, and there is no extra layer of clustering for the treatments. Thus, this design is *not* a PN-RCT because the design structure is parallel for treatments and controls. This design differs from the one in the previous bullet because it is answering a different question related to the relative effects of two interventions: Is small group instruction more effective than regular classroom instruction?

- Suppose now that the pull-out program *partially replaces* the normal classroom instruction. This could occur, for example, if the treatment group students are pulled out of their regular classrooms for part of time (for instance, for 15 minutes during a 1-hour reading instruction period). Although this design is a hybrid of the designs in the first two bullets above, it can be classified as a PN-RCT because of the extra layer of clustering in the treatment groups due to the ICs. However, because treatment and control students spend different amounts of time in their regular classrooms, the correlation structures due to shared regular classroom experiences could differ for the two groups of students.

- Suppose now that *classrooms rather than students* are randomly assigned to the treatment and control groups and that students in the treatment classrooms receive the pull-out intervention. In this design, classrooms are clusters, not blocks as in the previous examples. Thus, we have a C-RCT. However, if the pull-out program *supplements* or *partially replaces* the regular classroom instruction, we have classified these designs as "PN-RCTs" because there is an extra layer of clustering in the treatment group due to

the ICs. If the pull-out program instead *fully replaces* the traditional curriculum, then this design would not be a considered a PN-RCT because there is a parallel data structure for the treatment and control groups.

Although these designs may appear at first glance to be similar, they require quite different methods for estimating treatment effects (see chapters 3 and 4).

**What if teachers provide services in multiple ICs? Is this a PN-RCT?** The answer is "yes" in general, although the nesting structure becomes more complex. For example, consider an evaluation with 10 ICs and 2 teachers (Teachers A and B) where 5 ICs are taught by Teacher A and the other 5 ICs by Teacher B. In this case, treatment students in different ICs who are taught by Teacher A may have correlated outcomes because they are affected by the general competence of the teacher. In addition, treatment students in the same IC might be expected to have even greater similarity because they share not only the same teacher but also the same IC environment. Thus, we obtain a complex three-stage hierarchical structure for the treatment group (students nested within ICs nested within teacher), but no such clustering exists for the control group.

As another example of a design where teachers provide services in multiple ICs, consider an evaluation where a tutor has one-on-one sessions with treatment students but teaches several students in the sample. Although this design has only one student per tutoring session, it could be considered to be a PN-RCT if we assume that students with the same tutor are in the "same" IC. In this design, the intervention is administered to students individually, but students taught by the same tutor could have correlated outcomes, so this design could fit the PN-RCT structure.

**What confounding issues can arise in PN-RCTs that could bias the impact estimates?** Confounding occurs when an effect that is actually due to another source is attributed to the intervention. Confounding can lead to spurious impact findings to the extent that observed treatment-control differences can be explained by extraneous factors other than the offer of the treatment. Confounding could occur from a poorly designed study or from factors that occur after random assignment (such as treatment-control differences in study attrition). An example of confounding in education is an experiment in which exactly two teachers participate in the study: all of the students of Teacher A are assigned to the control group, and all of the students of Teacher B are assigned to the treatment group (see example 1.3 in chapter 1). In this example, the effect of the intervention cannot be separated from the effect of the teacher.

Extraneous variables that affect all students in the study equally are not confounding variables. For example, in an experiment conducted entirely in a high-poverty school district, poverty status is not a confounding variable because it affects both the treatment and control group students to an equivalent degree (although restricting the sample to a homogenous set of schools could reduce the generalizability of study findings to broader contexts). Rather, we are concerned about variables that systematically affect students in the treatment group differently from how they affect students in the control group.

PN-RCTs have several sources of potential confounding less commonly found in other RCTs. These originate because the treatment group has a different structure from the control group. A consideration of these confounding factors is an important dimension for assessing whether a PN-RCT design should be used for an evaluation. Examples of such confounding factors are

- *Grouping students into ICs can be considered to be an intervention in itself.* With a PN-RCT, the effect of the intervention being tested is *always* confounded with the effect of placing the treatment students in new types of clusters (without any intervention program). For instance, in a PN-RCT with a pull-out intervention that meets outside the normal classroom, it is possible that the treatment group students would experience performance changes by being placed in small groups *without* the pull-out curriculum. In this case, the observed treatment-control differences could reflect both grouping effects as well as the effects of the intervention curriculum. Grouping effects could lead to increased performance (for example, because of an improved learning environment or more attention from teachers) but could also lead to decreased performance (for example, because of stigma effects).

- *Student and teacher knowledge of the experiment might be greater in the treatment group.* Students in ICs are likely to know that they are participating in the study, especially for interventions that change the learning environment (e.g., a pull-out program). Control students may not have this same knowledge and may be less invested in evaluation outcomes than are treatment students. Thus, there may be a study effect, unrelated to the nature of the intervention, for students in merely being selected to receive the treatment.

- *Attrition may differ in the treatment and control groups.* If students in the control group are less involved in the study, they may be more likely to drop out before the outcomes are measured. In an alternative scenario, if an intervention requires large time commitments or is intrusive, students in the treatment group may be more likely to drop out. In either case, it is possible that the students with complete data for all outcomes may have different characteristics in the treatment and control groups. Nonresponse adjustments using available baseline data can be employed to help correct for differential attrition across the two research groups (for example, by adjusting the sample weights using propensity score procedures or using instrumental variable methods). However, if the reasons students drop out are related to the outcome variable

but cannot be fully explained by baseline covariates, then the differential attrition may lead the estimated treatment impact to be biased. Note that differential attrition between the treatment and control groups could also be an issue for I-RCTs and C-RCTs.

- *Sources and levels of measurement error may differ in the treatment and control groups.* Because students in the treatment group are in new types of clusters, it may be easier to administer the evaluation for these students (e.g., collecting data) than for control group students who may be more dispersed and less invested in the study. Furthermore, IC leaders may be more invested than control group educators in ensuring high study participation rates (e.g., obtaining signed study consent form from parents) and the collection of high-quality outcome data. In addition, students may perform differently if their teachers are supportive of the study than if they are not. A related issue is that outside data collectors may be more aware of (less "blinded" to) the research assignments of treatment students in the ICs than the control students, and this could affect data responses and quality. Treatment-control differences in data measurement and study participation could be related to study outcomes and intervention effects. These confounding factors could lead to biased impact estimates.

  In the BELL summer school study (example 1.3 in chapter 1), for example, some of the outcome measures were observed onsite at the BELL program. The control group students, however, were less likely to show up at the site for testing and, therefore, were more likely to have data collected later at their homes. Thus, the control students were, on average, tested later than the intervention students. The testing time was partially confounded with the intervention in the study.[8]

If a PN-RCT is to be conducted, care must be taken to help overcome these sources of confounding to ensure that the only difference between the two research groups after random assignment is the administration of the intervention. If confounding is likely to be a major concern in a PN-RCT, researchers may want to consider whether a C-RCT design could be implemented instead. For example, in some studies, it might be possible to have control students grouped in study halls of the same size as the ICs. Then both arms of the study would have clustering, and the study would measure whether the intervention has any effect beyond that observed by having students grouped together. Because the students are in a study hall, the outcome measurement can be administered more uniformly for the two study arms.

To help overcome these sources of confounding, education researchers could also consider allocating a greater share of evaluation resources to the control than treatment group for study recruitment and data collection efforts. This approach could help minimize attrition and data quality differences across the two research samples.

[8] Chaplin and Capizzano (2006) adjusted student test scores for the time lag between the program end and the test score measurement.

## 2.2 How Should ICs Be Formed?

A critical issue in designing a PN-RCT is the approach for assigning students and teachers to ICs to best answer the key evaluation research questions. In the education context, many factors can contribute to how students and teachers are placed into ICs. In some studies, students and teachers may be assigned randomly to ICs. In other studies, ICs may be formed naturally based on student characteristics (e.g., prior academic achievement and age), teacher characteristics (e.g., experience, preferences, and expertise), or practical constraints (e.g., students' daily schedules and tutor availability). In these cases, IC formation is not an evaluation design parameter but is part of the evaluation context. Regardless of how IC assignments are made, PN-RCTs are randomized experimental designs because individuals are randomly assigned to the treatment and control groups. Thus, PN-RCTs provide causal estimates of average treatment effects (full treatment-control differences) on key study outcomes. PN-RCTs are *not* quasi-experimental designs (QEDs), which are based on *comparison* groups that are formed using methods other than random assignment.

In some PN-RCTs, researchers may not have the flexibility to design the way ICs are formed; this might occur, for example, in evaluations of existing or ongoing interventions where educators have well-established mechanisms for forming the ICs. In other PN-RCTs, researchers may not want to alter the IC formation process (even if they could), so they can examine the effects of interventions as typically implemented; this approach could improve the generalizability of study results. In yet other evaluations, researchers may want to randomly assign students to ICs to address a broader set of research questions by comparing outcomes across ICs to rigorously assess the relative effects of particular IC features (for example, IC tutor characteristics).

This section discusses several design options for IC formation under PN-RCT designs, and the advantages and disadvantages of each one. The issue of how large an IC should be is discussed in section 2.4.

**Natural assignment of students and teachers to ICs.** For most PN-RCTs that have been conducted in the social sciences, study ICs were formed naturally after random assignment using IC assignment rules in place at the time of the evaluations. For example, in a large, nationwide random assignment evaluation of the Job Corps program—the nation's largest residential education and training program for disadvantaged youth ages 16 to 24—eligible program applicants were randomly assigned to treatment and control groups, and treatment students were then assigned to Job Corps

centers (the ICs) using existing program rules (Schochet et al. 2009). Similarly, in nearly all of the many experimental evaluations that have been conducted in the U.S. to examine the effects of group-based case management services for welfare recipients, unemployment insurance recipients, criminal offenders, drug addicts, and myriad other populations, ICs were formed using existing program rules. In these studies, researchers did not interfere with IC placements, so they could examine intervention effects in typical implementation settings.

The main advantage of a PN-RCT that uses status quo procedures for forming ICs is that the impact findings could be germane to the way in which the program would be rolled out more broadly. For instance, in an education PN-RCT, if students are normally tracked into ICs based on their ability level, there might be policy interest in conducting an evaluation that uses ICs formed in this manner rather than a design where students are randomly assigned to ICs, thereby undoing this tracking. Thus, the design with purposeful IC placements aims to address the following policy-relevant research question: What are the effects of the intervention as typically implemented?

There are, however, several disadvantages to this design. First, the design could reduce the precision of the impact estimates by increasing the estimated variances of outcomes across ICs (see section 2.6). Differences across ICs could arise due to systematic variation across ICs in the characteristics of students, the quality of teachers, the teacher-student-fit, the course curriculum, and the extent of peer interaction effects. If higher ability or lower ability students at baseline are systematically grouped together to facilitate group learning, the variance of outcomes across ICs could be sizable.

There are several ways to help mitigate this variation inflation problem. First, including student-level baseline covariates in the estimation models can improve precision by explaining some of the variation in student outcomes across ICs due to the nonrandom sorting of students to ICs (see section 3.4); this is the preferred method. Such covariates will vary depending on the evaluation context but could include measures of prior year academic achievement, risk factors for adverse outcomes (for example, special education or English language learner status), socioeconomic status, class schedules, and other indicators of how students were sorted into ICs. Thus, for PN-RCTs, it is important that researchers collect information on IC assignment rules and, to the extent possible, collect baseline data to measure these rules, including qualitative information from school staff. Second, ICs could be treated as fixed effects for the study rather than as randomly selected from a broader population, although such an approach may underestimate the true variability (see section 2.5). Finally, researchers could improve precision by selecting study sites with uniform types of students, learning environments, interventions, and providers. Although this approach will likely reduce IC heterogeneity, the cost is that the study results will likely generalize to a narrower

population. Thus, the desire to obtain more precision for the estimated treatment effects must be balanced with the desire to generalize to a wider population.

Alternatively, if ICs are formed based on student characteristics, a blocked PN-RCT design might be used, as described in section 1.2.2. In this design, the researcher would assign all participating students to blocks using the existing placement rules, where each block would have twice as many students as would typically be assigned to an IC. Then, within each block, half of the students would be randomly assigned to the treatment group and the other half to the control group. The students assigned to the treatment group within a block would form one IC and receive the intervention in the IC setting; the students in the control group in that block receive the control protocol. The benefit of this design is that information on IC placements would be available for *both* the treatment and control students, and block indicator variables could be added to the models. The blocks would control for all potential explicit and subjective factors that were used when assigning students to ICs. Thus, this blocked design could control for the mechanisms used to assign students to ICs, thereby increasing the precision of the impact estimates.

A second drawback of a PN-RCT where ICs are formed naturally is that the design would have no statistical basis for comparing the outcomes of treatment group students across ICs or comparing a subset of ICs to the control group. Thus, this design would not be suitable for addressing research questions such as: "Do intervention effects differ for more experienced tutors than less experienced ones?" Thus, under this design, it would be harder to rigorously assess the relative effects of particular IC features in improving student outcomes, which could be important to inform decisions about how best to target specific intervention services and how to improve the design or implementation of the tested interventions.

**Random assignment of students to ICs.** In some PN-RCTs, researchers may feel that the study research questions can best be addressed if students are randomly assigned to ICs. There are several advantages to this approach. First, the random assignment of students to ICs will, in expectation, balance student characteristics across ICs. Thus, this design could improve the precision of the average treatment effects (ATE) estimates by reducing variability across ICs, although these effects would still remain because of IC differences in teacher quality, peer effects, and sampling error. Statistical power issues might be an important consideration in smaller scale PN-RCTs that are conducted in only few schools or classrooms.

A second benefit of a design that randomly assigns treatment students to ICs is that it could be used to compare ICs to each other and to the control group. In particular, this design could be used to address research questions pertaining to the relative effectiveness of intervention providers with different characteristics, for example, teachers with varying experience or education levels, in improving students' academic achievement. This design could answer a research question such as: "What are intervention effects for the average student who is taught by an experienced teacher?" An example of an influential C-RCT that used this type of design was the Evaluation of Teach for America (TFA) (Decker et al. 2004) that compared TFA teachers with traditionally trained teachers. Students were randomly assigned to classrooms, so the evaluation could rigorously assess whether student achievement outcomes differed for the two types of teachers and across TFA teachers with different background characteristics.

If these types of questions are of particular importance for the evaluation, statistical power for these comparisons could be improved if a higher percentage of students were assigned to the treatment than control group. For example, if logistically feasible, rather than a design with a 50-50 treatment-control split, the design could use a 70-30 treatment-control split, so more ICs would be available for analysis. This approach, however, would reduce precision for the overall treatment-control contrast.

The main disadvantage of a design that randomly assigns treatment students to ICs is that the ATE estimates based on the full treatment and control groups might not pertain to intervention effects more generally. For example, consider a tutoring intervention where students are normally placed into small learning groups based on their reading ability, but where the evaluation instead randomly assigns students to ICs. In this case, it is likely that the IC learning environment for the study would not be typical of the IC learning environment if the intervention were rolled out more broadly. Thus, in this case, the study results might not be generalizable to broader settings, and policymakers could criticize study findings as not being germane to their own contexts. Clearly, this would not be a concern if, in typical intervention settings, students would be more or less randomly assigned to ICs. But the generalizability issue could be problematic if typical implementation involves the nonrandom tracking of students into ICs. This is an important concern because the intent of many evaluations is to identify promising interventions for broader use. Thus, to the extent possible, researchers may want to conduct evaluations in real world settings.

## 2.3 Can PN-RCTs Accommodate Multiple Treatment Groups?

In some PN-RCTs, education researchers might want to test the relative effects of several intervention components in isolation or in combination, so study results can be used to develop a promising package of intervention services. For example, an education researcher may be interested in varying intervention dosage levels across ICs (e.g., the number of tutoring sessions per week or session length) to examine the association between impacts and dosage. Or there may be interest in testing different intervention features (e.g., IC size and curriculum) in a factorial design. These designs have multiple treatment groups: in the first example, each dosage would be considered a different treatment, and in the second example, four treatment groups would be formed using the four possible combinations of IC size and curriculum.

These multi-armed PN-RCTs have the same general design structure as the PN-RCT with a single treatment and control group. The common feature of these designs is that the treatment students are nested within ICs. Thus, the statistical estimation framework discussed in chapters 3 and 4 can be easily generalized to the multi-armed design.

In some multi-armed experiments, there may not be a pure control group. In the PN-RCT context, this would mean that all students in the sample are assigned to ICs. This design is no longer partially nested but is fully nested because the same design structure applies to all research groups; this simplifies the estimation models.

For multi-armed PN-RCTs, it is preferable that students are randomly assigned to ICs along with the various treatment packages. This design is appropriate because multi-armed experiments are typically conducted to test out new treatments or combinations of treatments, and, thus, there might not be existing mechanisms for assigning students to these types of ICs. Furthermore, although in theory the multi-armed design could also be used for the design where ICs are formed naturally, results from this design would be difficult to interpret, and statistical power would likely be lower.

Researchers should realize that statistical power may be a concern in multi-armed PN-RCTs due to potentially small sample sizes in each treatment arm, as well as statistical adjustments to significance levels that should be used to account for the multiple comparisons problem (see Schochet 2009).

Thus, these designs will typically require large samples so that the study can rigorously compare the intervention packages to each other. These designs may not be suitable for small-scale PN-RCTs.

## 2.4 How Many ICs and Students Per IC Should Be Selected?

An important part of any evaluation design is determining appropriate sample sizes to ensure that the study will have a good chance of finding a statistically significant ATE estimate, if the true ATE is of a size that is meaningful and attainable. Lipsey et al. (2012) provide a framework for defining meaningful and attainable ATEs for education evaluations (where ATEs are measured in effect size or standard deviation units). For instance, they suggest that researchers examine the natural growth in student achievement in a school year, policy-relevant performance gaps across student subgroups or schools, and observed effect sizes from previous similar evaluations. Education researchers can use this framework for their evaluations to set targets for anticipated treatment effects—that is, "precision targets"—and to determine appropriate sample sizes to achieve those targets.

In education RCTs, the statistical precision of the impact estimates (often loosely referred to as the "statistical power" of the evaluation) depends on the variances of the ATE estimates (measured in effect size units). These variances are primarily a function of sample sizes and the design structure. For unclustered designs (I-RCTs), variances are primarily a function of the number of treatment and control students in the sample. For clustered designs (C-RCTs), variances are primarily a function of (1) the number of clusters, (2) the number of students per cluster, and (3) the correlation among two students in the same cluster—the intraclass correlation— described in chapter 3. For all designs, variances of ATE estimates can be reduced by including baseline covariates in the statistical model, and the reduction in variance can be measured using the regression $R^2$ value.

The basic PN-RCT design with random IC factors is a hybrid of an I-RCT and C-RCT design; thus, the power calculations for PN-RCTs must incorporate variance features of both designs. These power calculations will determine the required number of control group students, ICs, and treatment group students per IC so that the evaluation can estimate ATEs with the desired precision level.

In some PN-RCTs, the number of students per IC will be determined by program staff based on their normal IC assignment mechanisms. In these cases, researchers will need to assess the required number of ICs, given these IC student sizes. In other instances, researchers may have some latitude in determining IC student sizes for their evaluations. This choice could depend on the nature of the

intervention and its theory of change, implementation considerations (such as the availability of space and teachers), and precision and cost tradeoffs of selecting more ICs as compared to selecting more students per IC.

Chapter 4 provides a detailed mathematical discussion of sample size calculations and formulas for all the PN-RCT and related designs considered in this paper. Table 2 displays illustrative sample size results for the basic PN-RCT design with random IC effects.[9] The table displays total sample sizes—split evenly between the treatment and control groups—that are required to achieve ATE precision targets measured as minimum detectable impacts in effect size units (MDEs). The figures are presented for various intraclass correlations $(\rho_\theta)$, regression $R^2$ values that range from 0 (the model with no covariates) to .75, and IC sizes that range from 2 to 20 students to allow for designs with small ICs (e.g., groups with tutors) and larger ICs.

**Table 2. Total sample size calculations for students for the basic PN-RCT design with random IC effects, for treatment and control groups of equal size**

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Average IC sample size | 0 | .25 | .50 | .75 |
| MDE Target = .10; $\rho_\theta$ = .1 | | | | |
| 2 | 3,489 | 2,617 | 1,745 | 872 |
| 5 | 4,013 | 3,010 | 2,006 | 1,003 |
| 10 | 4,885 | 3,664 | 2,443 | 1,221 |
| 20 | 6,630 | 4,972 | 3,315 | 1,657 |
| MDE Target = .10; $\rho_\theta$ = .2 | | | | |
| 2 | 3,926 | 2,944 | 1,963 | 981 |
| 5 | 5,103 | 3,827 | 2,552 | 1,276 |
| 10 | 7,066 | 5,300 | 3,533 | 1,767 |
| 20 | 10,992 | 8,244 | 5,496 | 2,748 |
| MDE Target = .20; $\rho_\theta$ = .1 | | | | |
| 2 | 872 | 654 | 436 | 218 |
| 5 | 1,003 | 752 | 502 | 251 |
| 10 | 1,221 | 916 | 611 | 305 |
| 20 | 1,657 | 1,243 | 829 | 414 |
| MDE Target = .20; $\rho_\theta$ = .2 | | | | |
| 2 | 981 | 736 | 491 | 245 |
| 5 | 1,276 | 957 | 638 | 319 |
| 10 | 1,767 | 1,325 | 883 | 442 |
| 20 | 2,748 | 2,061 | 1,374 | 687 |

[9] Sample size calculations for the model where IC effects are treated as fixed are shown in section 4.3.3.

**Table 2. Total sample size calculations for students for the basic PN-RCT design with random IC effects, for treatment and control groups of equal size (Continued)**

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Average IC sample size | 0 | .25 | .50 | .75 |
| MDE Target = .30; $\rho_\theta$ = .1 | | | | |
| 2 | 388 | 291 | 194 | 97 |
| 5 | 446 | 334 | 223 | 111 |
| 10 | 543 | 407 | 271 | 136 |
| 20 | 737 | 552 | 368 | 184 |
| MDE Target = .30; $\rho_\theta$ = .2 | | | | |
| 2 | 436 | 327 | 218 | 109 |
| 5 | 567 | 425 | 284 | 142 |
| 10 | 785 | 589 | 393 | 196 |
| 20 | 1,221 | 916 | 611 | 305 |
| MDE Target = .40; $\rho_\theta$ = .1 | | | | |
| 2 | 218 | 164 | 109 | 55 |
| 5 | 251 | 188 | 125 | 63 |
| 10 | 305 | 229 | 153 | 76 |
| 20 | 414 | 311 | 207 | 104 |
| MDE Target = .40; $\rho_\theta$ = .2 | | | | |
| 2 | 245 | 184 | 123 | 61 |
| 5 | 319 | 239 | 159 | 80 |
| 10 | 442 | 331 | 221 | 110 |
| 20 | 687 | 515 | 343 | 172 |
| MDE Target = .50; $\rho_\theta$ = .1 | | | | |
| 2 | 140 | 105 | 70 | 35 |
| 5 | 161 | 120 | 80 | 40 |
| 10 | 195 | 147 | 98 | 49 |
| 20 | 265 | 199 | 133 | 66 |
| MDE Target = .50; $\rho_\theta$ = .2 | | | | |
| 2 | 157 | 118 | 79 | 39 |
| 5 | 204 | 153 | 102 | 51 |
| 10 | 283 | 212 | 141 | 71 |
| 20 | 440 | 330 | 220 | 110 |

**NOTES:** All calculations were conducted assuming a 5 percent significance level, two-tailed test at 80 percent power, and equal treatment and control group sample sizes. The figures in the table show total sample sizes (split evenly between the treatment and control groups) that are required to achieve the indicted ATE precision targets measured in effect size units. The figures are shown for various precision targets, intraclass correlations $(\rho_\theta)$ and regression $R^2$ values. See chapter 4 for formulas and details.

Our key results are as follows:

*In the basic PN-RCT design, accounting for the clustering of students in ICs will increase required sample sizes.* Consider the required total sample size of 957 students (split evenly between the treatment and control groups) that is displayed as the second shaded entry in the table.

This is the sample size that would be required if we anticipate that the ATE will be .20 standard deviations; the number of students per IC will be 5 (in 191 ICs); the intraclass correlation $\rho_\theta$ will be .10; and the regression $R^2$ value will be .25. Under an I-RCT design with no clustering due to the ICs, the required sample size would reduce to 589 students (not shown). Required sample sizes become smaller as $R^2$ and precision targets increase and as $\rho_\theta$ values decrease. $R^2$ values have a particularly large effect on precision; thus, the collection of detailed baseline variables is an important design feature for improving the precision of the impact estimates.

***Precision levels can typically be improved if more ICs and fewer students per IC are sampled for the study.*** In the above example, using the shaded figures in Table 2, we find that reducing the IC sample size from 5 to 2 students will reduce the required sample size from 957 students (in 191 ICs) to 736 students (in 368 ICs). Thus, subject to study resource and implementation constraints, researchers designing PN-RCTs should be aware that statistical power can be improved by selecting more ICs with fewer students per IC. This design could be implemented, for example, by subsampling students within ICs to allow for a larger sample of ICs. The subsampling should be done randomly to ensure that the selected sample is representative of all students in the ICs.

We refer interested readers to chapter 4 for the sample size formulas that were used to calculate these quantities and those for other PN-RCT designs considered in chapters 3 and 4. These formulas can be used by researchers to perform power calculations for their own evaluations using specific power analysis parameter values that fit their contexts.

## 2.5 What Should Be Done About Other Sources of Clustering in PN-RCTs?

In any experimental evaluation of an education intervention, students in the sample are likely to be connected in some way that could lead to correlations among their outcomes. Regardless of the study design, students may be clustered in classrooms, schools, or neighborhoods. Students eligible for the study could also have social or familial connections to other students.

Because sources of clustering exist in all evaluations, as background for discussing statistical analysis for PN-RCTs, it is useful to ask the following general question that pertains to all experimental

designs: When is it necessary to statistically account for clustering effects in the analysis? This question has caused quite a bit of confusion in the literature.

*In general, only clustering that occurs as part of the design of the experiment needs to be considered in the analysis.* Other forms of clustering do not need to be considered. This idea seems somewhat non-intuitive at first. Why would one type of clustering be considered in the analysis but not another? The answer lies in the randomization process.

Let us first consider this issue for an I-RCT with a pool of students from different schools. It is almost impossible that none of the students have connections to others. Some of those connections may be observed (such as attending the same school), while others are unobserved (such as belonging to the same church, softball team, or play group). When those students are individually randomly assigned to treatment or control groups, however, the randomization process randomly divides students who originate in the same school to the two research groups. The randomization essentially cancels out the pre-existing clustering effect from the original schools, just as it cancels out pre-existing effects from unobserved connections between the students such as belonging to the same church, softball team, or play group. If the randomization were redone, a different subset of students from a specific school or church or softball team or play group would end up in the treatment group. Thus, in the analysis of data from an I-RCT design, the models must account for the variance of outcomes across students (because students are randomized) but not across the pre-existing clusters (which cancel).

The situation is different in a C-RCT. There, the pre-existing *clusters* are the units that are randomized to the treatment or control group. Let's suppose a C-RCT is conducted in which the clusters are schools. Students in the same school always end up together in the same study arm. If the school is randomized to receive the intervention, all the students in the school receive the intervention; if the school is randomized to receive the control, all the students in the school receive the control. If the randomization were redone, the students in the same school would still end up together. Thus, for a C-RCT, the analysis must account for the variability of outcomes across clusters (that we label as "clustering effects"). There is widespread agreement across many disciplines that clustering effects due to the random assignment of clusters must be taken into account in the analysis (see, for example, Donner and Klar 2000, 2004; Murray 1998; Raudenbush 1997; Schochet 2008; Walsh 1947). In a C-RCT, the random assignment of clusters to treatment or control provides the basis for inference; an analysis that ignores the clustering effects will calculate "standard errors" that are misleadingly smaller than the true standard errors. As Cornfield (1978,

p. 101) said, "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception."

In a PN-RCT in which individual students are randomly assigned to the treatment or control group, the randomization essentially cancels out the effects of pre-existing clustering, just as in an I-RCT. The randomization does not, however, cancel out the effect of clustering by the ICs. If the experiment were repeated and new students were randomized to the treatment and control groups, the students in the treatment group in the replicated experiment would also experience clustering from the ICs. The question then becomes how the analysis should treat clustering effects induced by the formation of the ICs *after* randomization. This is a design *choice* based on researcher assumptions. This issue gets at the heart of the fixed-versus-random IC effects issue that we discuss next.

## 2.6 An Overview of the Literature on Statistical Analysis of PN-RCTs: How Should IC-Level Clustering Be Treated?

The statistical analysis of the basic PN-RCT design has been considered in other disciplines, particularly in the area of psychological counseling. Martindale (1978) discussed experiments where control group subjects receive no psychotherapy while treatment group subjects receive psychotherapy, with multiple subjects being treated by the same therapist. Although the intervention is not conducted in a group setting, patients seeing the same therapist are expected to have more similar outcomes than patients who see different therapists; essentially, the group of patients in the study seeing one therapist forms an IC.

Since Martindale's (1978) article, there has been a robust debate in the literature on the proper statistical analysis of PN-RCTs for experimental evaluations of psychotherapy interventions (Crits-Christoph and Mintz 1991; Crits-Christoph, Tu, and Gallop 2003; Hoover 2002; Serlin, Wampold, and Levin 2003; Siemer and Joormann 2003; Wampold and Serlin 2000). The central issue addressed by this literature is how "therapist effects" or, in our terminology IC effects, should be handled in the impact analysis. The statistical problem is that even though random assignment is conducted at the patient level, treatment patients are clustered (nested) within therapists and, thus, may have correlated outcomes that could inflate the variance estimates.[10]

[10] More precisely, these studies consider designs where patients are randomly assigned to one of two treatment groups that are offered different types of therapy. Thus, therapist effects would apply to both research groups in these designs. The statistical issues associated with these designs are similar to those discussed in this report.

The debated issue for these designs is whether IC effects should be treated as (1) *random* factors where the model error structure includes random IC effects or (2) *fixed* factors where the model excludes the random IC effects. The central question is whether one should regard intervention providers as a random factor for purposes of generalizing the impact findings to a broader provider population or whether one should regard providers as a fixed factor, thereby restricting inferences to the particular providers used in the study. Several articles in the 2003 issue of *Psychological Methods* (volume 8) take differing stances on this issue.

In the education context, under the random effects assumption, ICs are assumed to be sampled from a larger population of possible IC realizations. It is assumed that in a re-run of the experiment, not only would the specific students assigned to the treatment and control groups change (because of random assignment), but such factors as IC assignment mechanisms, intervention teachers, curricula, and implementation fidelity could also change. Thus, under this approach, researchers would use an HLM approach for estimation where ICs are treated as another HLM level. Accordingly, variances of the estimated ATEs would be inflated to account for the extent to which mean outcomes vary across ICs. In this framework, the study results generalize outside the study sample (see, for example, Baldwin et al. 2005; Bauer et al. 2008; Lee and Thompson 2005; Roberts and Roberts 2005; Sanders 2011; Siemer and Joorman 2003; Wampold and Serlin 2000).

Under the fixed effects assumption, ICs are instead assumed to be fixed (not sampled) for the study. This approach assumes that study findings pertain only to the IC assignment mechanisms, specific teachers, curricula, and service delivery that were in place at the time of the evaluation. Under this framework, it is assumed that in a re-running of the experiment, the treatment and control students would change but that the IC structure would remain fixed (that is, the same tutors and curriculum would be in place, and the same procedures would be used to assign students and tutors to the ICs). In this scenario, variability between ICs is soaked up in the student-level variance. In essence, with student-level random assignment, this approach reduces to an I-RCT, and simple ordinary least squares (OLS) methods can be used for estimation. Siemer and Joorman (2003) argued that if therapists are deliberately chosen for the study, they may not be representative of a larger population of therapists, and, therefore, a fixed effects approach should be considered. In that case, the results of the study apply to the treatment as delivered by the specific IC providers in the particular study.

In education evaluations, some IC-level factors are likely to be fixed, and some are likely to be random. Thus, the decision of whether education researchers should adopt the random or fixed framework is complex. To help navigate this choice, it is helpful to categorize sources of variation across ICs into two groups: (1) *student-sorting factors* due to pre-existing differences in the baseline

characteristics of students assigned to ICs; and (2) *intervention-related factors* due to differences across ICs in the characteristics of IC tutors, instructional approaches, peer effects, and any other unmeasured factors that might affect all of the students in an IC.

***Variation due to student-sorting factors.*** Under the random effects specification, it is important that the variance estimates not reflect pre-existing differences in the baseline characteristics of students across ICs that are correlated with key study outcomes. If students are tracked into ICs based on these characteristics, the estimate of IC-level variability under the random effects specification will in general overestimate the true IC-level contribution to the variance because it will capture student compositional differences across ICs. This could lead to variance estimates in the treatment group that are unrealistically large. For example, suppose that there is considerable tracking of students into ICs and that there are zero treatment effects for all sample members. In this case, the random effects framework will generate a much larger estimated variance for the treatment group mean than the control group mean simply because of the nonrandom sorting of students into ICs.

We, therefore, recommend that if the random effects approach is adopted and students are nonrandomly sorted into ICs based on their characteristics, *detailed baseline data should be collected for modeling the student assignment to ICs.* Model covariates are needed to explain differences in student compositional differences across ICs. These covariates must be available for *both the treatment and control students*. The idea is that after conditioning on the covariates, the model will mimic the design where students are randomly assigned to ICs. The exact covariates to include in the model will depend on the specific IC formation process but are likely to include pre-intervention achievement measures (such as pre-test scores and grades) and demographic variables (such as language ability and grade level). If such baseline data are not available, researchers could consider analyzing the data using the fixed effects framework, but this approach will likely understate the true variance and is not the recommended approach.[11]

***Variation due to intervention-related factors.*** Absent student-sorting factors, variability across ICs will primarily reflect intervention-related factors, such as differences in teacher quality, IC curriculum, peer effects, and so on. This type of IC-level variation may be of policy interest because it could reflect variation that might be expected in more widespread implementation of the

[11] We define the fixed effects model (with student-level randomization) as a standard I-RCT or OLS impact model that (1) does *not* include random IC factors in the model error terms and (2) does *not* include indicators of IC membership as model covariates for the treatment group. If students are tracked into ICs, including these IC indicators in the model may underestimate the variance of the treatment group mean because these IC fixed effects will attribute student-level characteristics to the ICs for the treatment group but not for the control group; thus, these indicators will remove too much of the student-level variability from the variance of the estimated impact (Lockwood et al., 2013).

intervention. Accordingly, education researchers may wish to adopt the random effects specification to account for this type of IC-level variation.

A complication, however, is that the variation in ICs due to intervention-related factors could have both fixed components (for example, tutor experience) as well as random components (for example, teacher-student interactions and peer effects). If researchers deem that the fixed components are important, the random effects specification may produce variance estimates that are upwardly biased. One approach for handling this issue would be to include fixed IC-level factors (for example, teacher experience) as model covariates (see section 3.4). This scenario might be germane, for instance, if schools targeted for the intervention always have a mix of experienced and inexperienced teachers, in which case it might be of policy interest to assess IC-level variation *conditional* on teacher experience (that is, across teachers with the same experience level).

***Summary and focus of paper.*** In education evaluations, some IC-level factors are likely to be fixed and some are likely to be random. Thus, it is likely that the random effects framework will provide upper bound estimates on the "true" variances for the treatment group, while the fixed effects framework will likely provide lower bound estimates. The main concern for the random effects framework is that the estimation models adequately control for the potential variation across ICs due to the nonrandom sorting of students to ICs. This type of "spurious" variation could seriously reduce the precision of the impact estimates and erode statistical power. The IC-level variance due to intervention-related factors is of greater policy interest.

In the remainder of this paper, we assume the random effects framework with rich baseline covariates available to explain some of the variation across ICs in student composition. This approach provides a unified framework for estimating ATE effects and their standard errors that account for the heterogeneity of treatment effects that could exist across the student population.

This approach is also appealing in that it likely provides conservative upper bound estimates on the "true" variances of the impact estimates. Finally, the random effects approach is more general than the fixed effects approach because it reduces to the fixed effects approach if the IC-level random effect is removed from the model error term (or equivalently, if the IC random effect is assumed to have zero variance; see chapter 3). Thus, the models that we discuss in this paper largely apply also to the fixed effects framework, with appropriate model restrictions.

## 2.7 What Are Key Data Collection Issues to Help Interpret the Study Findings?

There are several key data collection issues that are particularly important for PN-RCT designs. First, evaluators should collect identifying information on the ICs that each treatment group member attended during the follow-up period, including dates of attendance and changes in IC membership status over time. These data will be necessary to describe IC characteristics and IC formation. These data will also be needed to create covariates to include in the impact estimation models to explain some of the variation in mean outcomes across ICs, thereby increasing the precision of the estimates (as discussed more formally in chapter 3).

Second, for PN-RCT designs, it will be important for the evaluation to collect process information on how interventions and providers were selected for the study. Data should also be collected on key IC features, such as student and teacher characteristics, teacher practices, IC curricula, the fidelity of intervention implementation within each IC, and the mechanisms used to assign students and teachers to ICs. These process data can be collected as part of site visits, from teacher or principal interviews, or from other sources. These data can be used to form model covariates to adjust for pre-existing differences in the baseline characteristics of students and teachers assigned to ICs. The analysis of these data will also provide contextual information to help explain any observed variation in outcomes across ICs and the credibility of the random or fixed effects paradigm. This information can be used to help interpret the overall and subgroup impact findings and the extent to which they can be replicated and generalized to broader settings.

## 2.8 Do PN-RCT Issues Apply Also to Quasi-Experimental Designs?

Thus far, we have considered partially nested designs in experimental settings only. It is not always possible, however, to conduct RCTs of education interventions for a variety of logistical and ethical reasons. Instead, education researchers often conduct evaluations using QEDs such as matched comparison group, RD, or IV designs. Although a full consideration of partially nested designs in QED settings is beyond the scope of this paper, we briefly comment on a few big picture issues.

First, similar to experiments, QEDs all involve the identification of a treatment group that receives intervention services and a comparison group that does not. Furthermore, in all partially nested designs, the treatment group receives intervention services in clusters, whereas IC-level clustering does not exist in the control group. Thus, the design issues that we have discussed for PN-RCTs,

such as IC formation for the treatment group, the handling of IC clustering in the analysis, and data needs are largely germane to QEDs also.

Second, in QEDs where treatment group members are matched to comparison group members from another data source (using propensity score matching or related methods), overall statistical issues for impact estimation should be very similar to those for PN-RCTs. In these QEDs, the comparison sample serves as a proxy for an experimental control group; thus, similar differences-in-means estimation methods apply. This will not be the case, however, for other QEDs (such as RD and IV designs) that involve alternative impact estimation strategies to account for sample selection biases (e.g., local linear methods for estimating local average treatment effects for RD designs and two-stage least squares methods for IV designs). These methods have been adapted for C-RCT designs; thus, we surmise that they can also be adapted to partially nested designs.

Finally, QEDs typically require substantially larger samples than RCTs to achieve the same level of statistical precision. For instance, RD and IV designs require at least three to four times as many students as RCTs (Schochet 2009, 2011). This occurs in RD designs because of the substantial correlation (by design) between two model covariates: (1) the treatment status indicator variable and (2) the score variable used to determine who receives the treatment. In an IV design, precision is a function of the correlation between the instrument and treatment status, which is often low. Thus, education researchers designing partially nested QEDs will need to conduct careful power analyses to ensure that their evaluations have sufficient samples to produce precise impact estimates. Similarly, they will need to think carefully about the feasibility of conducting exploratory analyses (e.g., comparing outcomes across ICs) and multi-armed evaluations.

# Statistical Analysis of the Basic and Blocked PN-RCT Designs

# 3

This chapter discusses statistical methods that can be used to analyze data from a basic or blocked PN-RCT where IC factors are treated as random. We first consider the simplest PN-RCT design, where students from one sample or population—such as a single school, school district, Head Start center, or summer program location—are randomly assigned to experimental conditions. It is assumed that treatment group students are subsequently placed into ICs, whereas control group students are not. We consider IC effects but not clustering effects due to other factors, such as the schools and classrooms that students regularly attend. Thus, in this design, the clustering of treatment group students in ICs is the only source of clustering that might occur in the data set. As discussed in chapter 2, we assume that IC effects are random. We consider analysis of a continuous post-intervention measure, such as an achievement test score, that is analyzed at the student level.

We start with the simplest PN-RCT design in this chapter to illustrate basic concepts of impact estimation for PN-RCT designs. Chapter 4 presents estimation approaches for more complex models, such as hybrid C-RCT designs, where schools or classrooms are randomized, and treatment students are subsequently placed into ICs, and cross-nested designs.

The complicating feature of a basic PN-RCT is that the treatment group and the control group have different data structures. The treatment group has clustering, while the control group does not. As background for showing how these features can be unified in the basic PN-RCT, in section 3.1 we first review models for estimating treatment effects in I-RCTs and C-RCTs. Section 3.2 then provides an overview of the estimation theory for the basic PN-RCT design.

In section 3.3, we provide a constructed example with SAS code[12] and output that illustrates how to analyze the data from the basic PN-RCT design. Section 3.4 describes how to include covariates in the model. Section 3.5 then presents the model and shows how to estimate treatment impact from a blocked PN-RCT. The mathematical theory for these models is given in appendix A.

[12] Version 9.3 of SAS software was used to develop all of the code presented in this report. The code uses some features that are not available in earlier versions of SAS.

## 3.1 Review: Statistical Models for I-RCTs and C-RCTs

**Statistical Model for an I-RCT**. In an I-RCT, all observations can be treated as being independent. There are $n_T$ students in the treatment group and $n_C$ students in the control group, for a total of $n = n_T + n_C$ students in the experiment. Let $y_j$ represent an outcome score for student $j$. In the control group, assume that

$$y_j = \beta_0 + \varepsilon_j,$$

where $\beta_0$ is the theoretical mean of the control group and $\varepsilon_j$ is a normally distributed random variable with mean 0 and variance $\sigma^2$. In the treatment group, it is assumed that the observations have the same variance, but the treatment mean can differ. Let $\beta_0 + \beta_1$ denote the theoretical mean of the treatment group, so the ATE is given by $\beta_1$. Then the model for an observation in the treatment group is

$$y_j = \beta_0 + \beta_1 + \varepsilon_j,$$

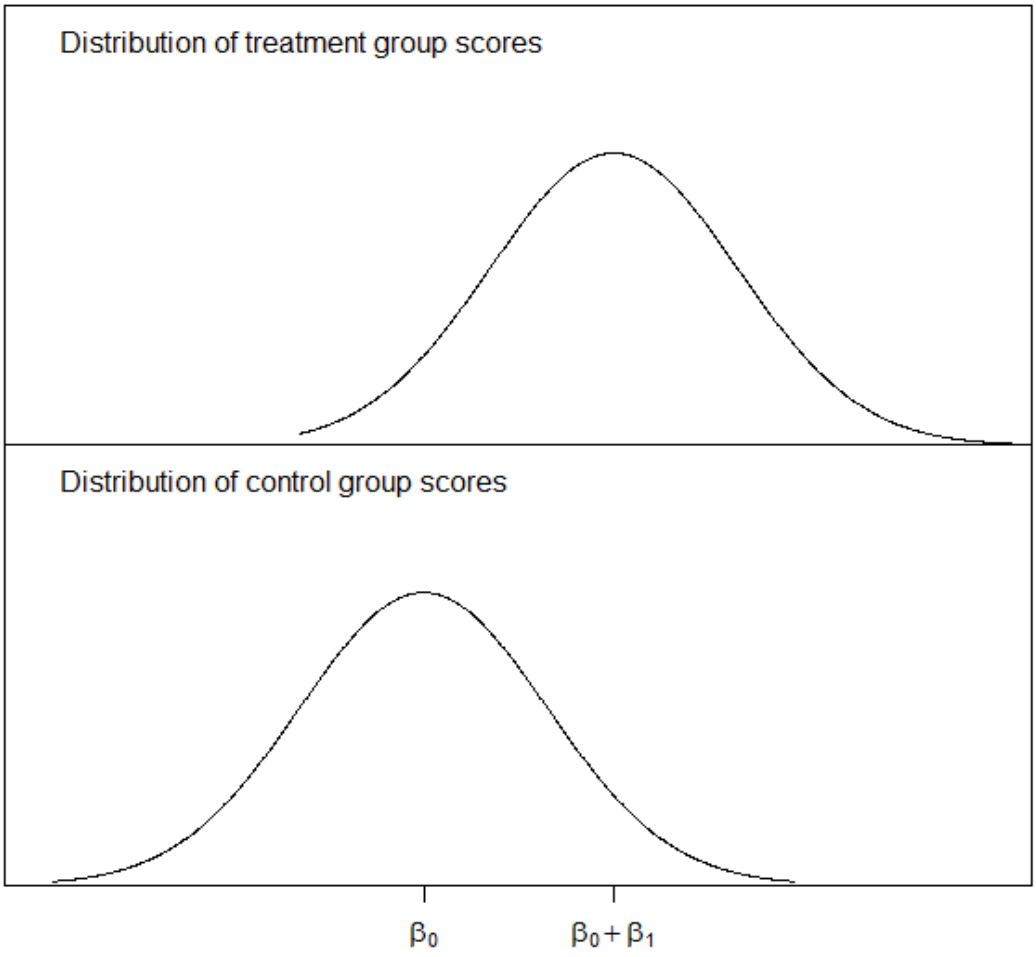
where again $\varepsilon_j$ is a normally distributed random variable with mean 0 and variance $\sigma^2$. In this model, we assume that the variance of the students in the control group is the same as the variance of the students in the treatment group.

We can present the model in unified form by using an indicator variable $T_j$ that describes whether the student is in the treatment group or the control group. Let $T_j = 1$ if student $j$ is in the treatment group, and $T_j = 0$ if student $j$ is in the control group. The statistical model may then be written in a unified form as

$$y_j = \beta_0 + \beta_1 T_j + \varepsilon_j, \tag{3.1}$$

where $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$. Because the experiment is randomized at the individual level, the error terms $\varepsilon_j$ can be assumed to be independent random variables. Figure 4 displays the distribution of student test scores for the treatment and control groups under the model in equation (3.1).

Figure 4. Distribution of test scores for treatment and control groups for an I-RCT.




Under this model, the least squares estimate of the ATE $\beta_1$ is

$$\hat{\beta}_1 = \bar{y}_T - \bar{y}_C,$$

where $\bar{y}_T$ is the mean score of the treatment group students, and $\bar{y}_C$ is the mean score of the control group students. *Under the I-RCT design, which allows us to assume that all students are independent,* the variance of the ATE is

$$Var_{I-RCT}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{n_T} + \frac{\sigma_\varepsilon^2}{n_C}.$$

The variance of the sample mean of each group is $\sigma_\varepsilon^2$ divided by the sample size; this is the formula used for the two-sample *t* test as taught in introductory statistics classes.

Note that in some experiments, the student-level variances may differ in the treatment and control groups. Then, instead of assuming that $\sigma_\varepsilon^2$ is the same for both groups, we may generalize the model to assume that $\varepsilon_j \sim N(0, \sigma_{\varepsilon C}^2)$ for observations in the control group and $\varepsilon_j \sim N(0, \sigma_{\varepsilon T}^2)$ for observations in the treatment group. Then,

$$Var_{I-RCT}(\hat{\beta}_1) = \frac{\sigma_{\varepsilon T}^2}{n_T} + \frac{\sigma_{\varepsilon C}^2}{n_C}.$$

**Statistical Model for a C-RCT**. C-RCTs are designs in which random assignment to the treatment and control groups is conducted at the cluster level. These designs are generally used for educational studies when the intervention affects all students in a classroom (such as a teacher professional development program) or in the entire school (such as a schoolwide behavior modification initiative). In education experiments, the clusters are usually classes, schools, or school districts, although sometimes families or cities serve as clusters. Suppose that $H$ classes are available for the experiment and that $I_T$ of the classes are randomly assigned to the treatment group, and the remaining $I_C = H - I_T$ classes are in the control group. Students in the same class share a teacher, classroom environment, and other factors that cause their outcomes to be more similar than outcomes of students in different classes. Let

$y_{ij}$ = test score of student $j$ in cluster $i$
$T_i$ = 1 if cluster $i$ is in the treatment group, 0 if in the control group
$\theta_i$ = random effect of cluster $i$
$\varepsilon_{ij}$ = student-level error (residual) for student $j$ in cluster $i$.

The observed outcome for student $j$ of cluster $i$ is $y_{ij}$; this is indexed by a double subscript to keep track of the cluster membership as well as the student within the cluster. The treatment indicator $T_i$ depends only on the cluster: if a cluster is assigned to the treatment group, then all of the students in that cluster will receive the treatment protocol.

An HLM is often used to account for the effects of the clustering in a C-RCT. The HLM captures that additional similarity by including an extra error term in the model. The *random effect* $\theta_i$ is a

random variable that captures variability at the cluster level that cannot be explained by the baseline covariates included in the model. The same random effect applies to all students in the same cluster, which is how the model accounts for correlated outcomes. Clustering occurs for both the treatment and control groups. The random effect is assumed to follow a normal distribution with mean 0 and variance $\sigma_\theta^2$; the larger the value of $\sigma_\theta^2$, the larger the effects of teacher, classroom environment, and other class-level factors.

The model for the response of a student in a C-RCT is

$$y_{ij} = \beta_0 + \beta_1 T_i + \theta_i + \varepsilon_{ij}, \tag{3.2}$$

where $\theta_i \sim N(0, \sigma_\varepsilon^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, and all of the variables $\theta_i$ and $\varepsilon_{ij}$ are independent. The random effect $\theta_i$ in equation (3.2) captures unmodeled factors at the cluster level. For example, suppose that clusters are classrooms and that class $i$ has an excellent teacher. Then, $\theta_i$ is likely to be positive, and that positive quantity is added to the scores of all the students in that class. Because each student in class $i$ shares the same value of $\theta_i$, students in the same class tend to be more similar than students in different classes; if $\theta_i$ is positive, all of the students start with a tendency to be above the mean for the treatment group.

If each class has the same number of students, the generalized least squares estimate of the ATE $\beta_1$ in a C-RCT is the same as in an I-RCT:[13]

$$\hat{\beta}_1 = \bar{y}_T - \bar{y}_C,$$

where $\bar{y}_T$ is the mean score of the treatment group students, and $\bar{y}_C$ is the mean score of the control group students. Because of the clustering, however, the variance of the ATE is larger in a C-RCT than in an I-RCT. If all classes have the same number of students and there are $I_T$ classes in the treatment group and $I_C$ classes in the control group, then

$$Var_{C-RCT}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{n_T} + \frac{\sigma_\theta^2}{I_T} + \frac{\sigma_\varepsilon^2}{n_C} + \frac{\sigma_\theta^2}{I_C}.$$

[13] If the classes have different numbers of students, the ATE is calculated by subtracting a weighted average of the control cluster means from a weighted average of the treatment cluster means, where the weights depend on the sample sizes and the estimated variance components; see appendix A.

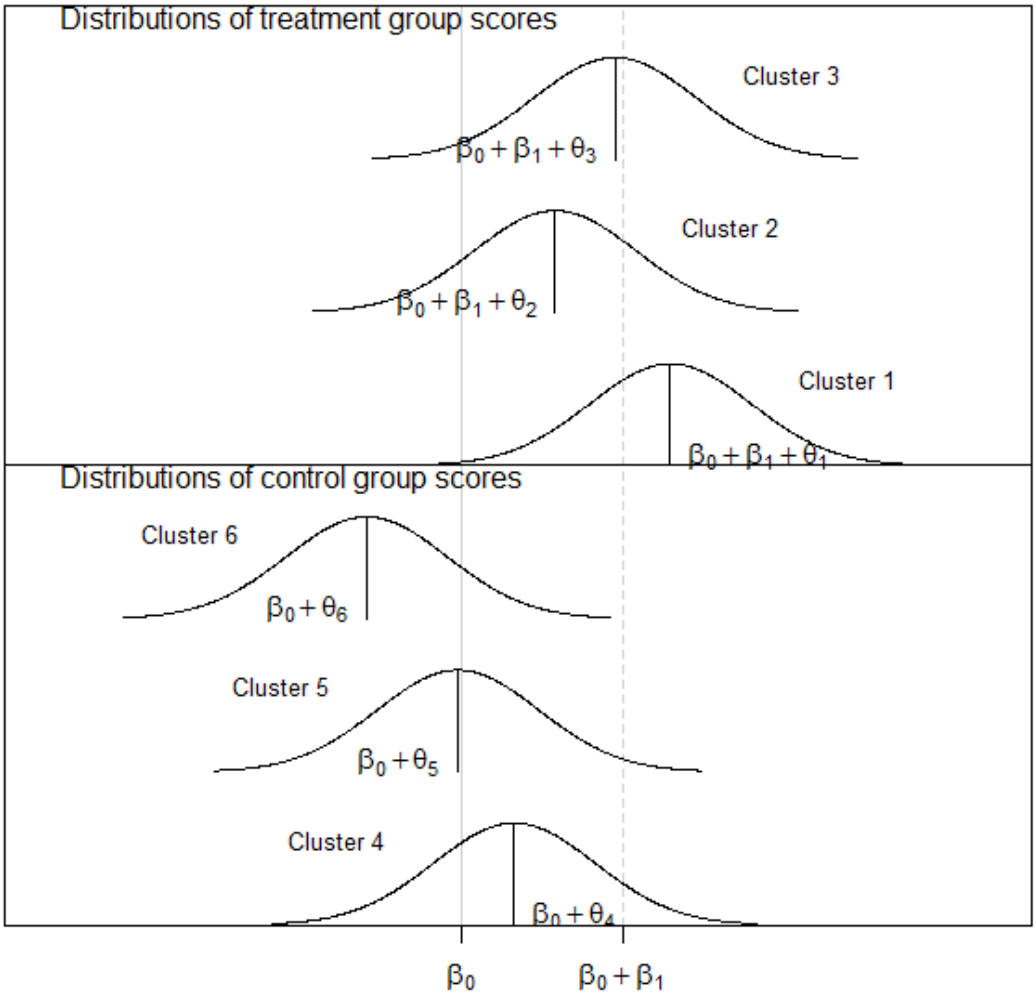
Note that $Var_{C-RCT}(\hat{\beta}_1)$ has two additional terms not found in $Var_{I-RCT}(\hat{\beta}_1)$, due to the presence of the clustering terms: $\sigma_\theta^2 / I_T$ and $\sigma_\theta^2 / I_C$. The only way to decrease those terms in the variance of the ATE is to increase the number of clusters.

Figure 5 shows a possible distribution of student test scores in the treatment and control groups under a C-RCT design in which $\sigma_\theta^2 = \sigma_\varepsilon^2 / 2$. Each of the six clusters has its own mean score, and the students in that cluster are more similar to each other because they share that same value for $\theta_i$ (that is, a common cluster mean). The intraclass correlation coefficient (ICC) for students in the same cluster is

$$\rho = \frac{Cov\left(y_{ij}, y_{ik}\right)}{\sqrt{Var\left(y_{ij}\right)Var\left(y_{ik}\right)}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}.$$

The ICC measures the degree of similarity among students in the same cluster. For the situation displayed in figure 5, $\rho = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} = 1/3$. If $\rho = 0$, then the clusters induce no similarity, and all observations are essentially independent. If we were to redraw figure 5 for the case of $\rho = 0$, all of the normal distributions in the control group would have mean $\beta_0$, and all of the normal distributions in the treatment group would have mean $\beta_0 + \beta_1$. If $\rho = 1$, then all of the students in the same cluster have the same test score. If we were to redraw figure 5 for the case of $\rho = 1$, the individual cluster normal distributions would have no variability, and the only variance would come from the cluster means. If $\rho = 1$, there is no advantage to observing more than one student per cluster because they all provide the same information. Typically, in education studies, the ICC when clusters are schools is in the range of 0.05 to 0.25 (Hedges and Hedberg 2007; Schochet 2008).

Figure 5. Distribution of test scores for treatment and control groups for a C-RCT.




As with the I-RCT, the model for a C-RCT can be generalized to allow for different cluster-level and different student-level variances in the treatment and control groups. To allow for different variances in the treatment and control groups, we can assume that $\theta_i \sim N\left(0,\sigma^2_{\theta C}\right)$ and $\varepsilon_{ij} \sim N(0,\sigma^2_{\varepsilon C})$ for clusters in the control group, that $\theta_i \sim N\left(0,\sigma^2_{\theta T}\right)$ and $\varepsilon_{ij} \sim N(0,\sigma^2_{\varepsilon T})$ for clusters in the treatment group, and, as before, that all of the variables $\theta_i$ and $\varepsilon_{ij}$ are independent. Such a model might be appropriate if the intervention is expected to result in greater similarity among the students in a cluster. If, for example, the classes randomized to the control group have a study hall while classes randomized to the treatment group have different volunteers come in to

teach students about nutrition, we might expect the variances of the treatment and control group to differ at both the student and class levels. The control group would have the usual class-to-class variability and within-class variability of $\sigma^2_{\theta C}$ and $\sigma^2_{\varepsilon C}$. The treatment group, however, might have $\sigma^2_{\theta T} > \sigma^2_{\theta C}$ because the volunteers might have very different levels of effectiveness, and the class means in the treatment group would be more variable than those in the control group. Alternatively, or additionally, the within-class variability of the treatment students, $\sigma^2_{\varepsilon T}$, might be less than that of the control students, $\sigma^2_{\varepsilon C}$, if the treatment results in more cohesion among students in the same class.

## 3.2 Statistical Model for the Basic PN-RCT

The basic PN-RCT model with random IC effects is a hybrid of the I-RCT and C-RCT models. The students in the treatment group follow the C-RCT model, while the students in the control group follow the I-RCT model. Suppose that there are $I_T$ ICs in the treatment group and that there are a total of $n_T$ students in the treatment group and $n_C$ students in the control group.

Let's start with the model for the treatment group. The treatment group has clustering, so its model follows the form in a C-RCT. Let $y_{ij}$ denote the test score of student $j$ from IC $i$, for $i = 1$ to $I_T$, just as in the C-RCT. The subscript $j$ refers to students, with $j = 1$ to $J_i$ for treatment group students in IC $i$. The total number of treatment students is $n_T = \sum_{i=1}^{I_T} J_i$. The model for a test score in the treatment group is

$$y_{ij} = \beta_0 + \beta_1 + \theta_i + \varepsilon_{ij},$$

where $\beta_0 + \beta_1$ is the mean score for students in the treatment group, $\theta_i \sim N(0, \sigma^2_\theta)$ is a random effect for IC $i$, and $\varepsilon_{ij} \sim N(0, \sigma^2_{\varepsilon T})$ is the student-level error term for student $j$ in IC $i$.

The control students are not formed in ICs. To express the model in unified fashion, we let $y_{0j}$ denote the test score of student $j$ in the control group (the "0" denotes that the student does not belong to an IC). Students in the control group have no variability at the IC level, so the test score of student $j$ in the control group can be modeled as

$$y_{0j} = \beta_0 + \varepsilon_{0j}.$$

Here, $\beta_0$ is the mean score for students in the control group, and we assume that $\varepsilon_{0j} \sim N(0, \sigma^2_{\varepsilon C})$ is the student-level error term for student $j$ in the control group, for $j = 1$ to $n_C$. Note that for flexibility in modeling, we allow the student-level variability in the control group, $\sigma^2_{\varepsilon C}$, to differ from the student-level variability in the treatment group, $\sigma^2_{\varepsilon T}$. The student-level variances could differ for the treatment and control groups because $\sigma^2_{\varepsilon T}$ reflects variation *within ICs* for the treatment group, whereas $\sigma^2_{\varepsilon C}$ reflects variation across the *entire* control group for control students. It is possible that peer interactions or the intervention could result in more homogeneity within an IC for the treatment group, so the model allows for that possibility. When fitting models in section 3.3, we allow the student-level variance to differ in the treatment and control groups.

The separate models for the control and treatment groups can be combined into one unified model (this will be important in section 3.4, when we include other covariates in the model) by defining the treatment status indicator variable $T_{ij} = 1$ if student $j$ in IC $i$ is in the treatment group and 0 otherwise. Note that because ICs are formed only in the treatment group, $T_{ij} = 1$ for all students when $i = 1$ to $I_T$, and $T_{0j} = 0$. Then, for $i = 0, 1, \ldots, I_T$, we can write

$$y_{ij} = \beta_0 + \beta_1 T_{ij} + \theta_i T_{ij} + \varepsilon_{ij}. \tag{3.3}$$

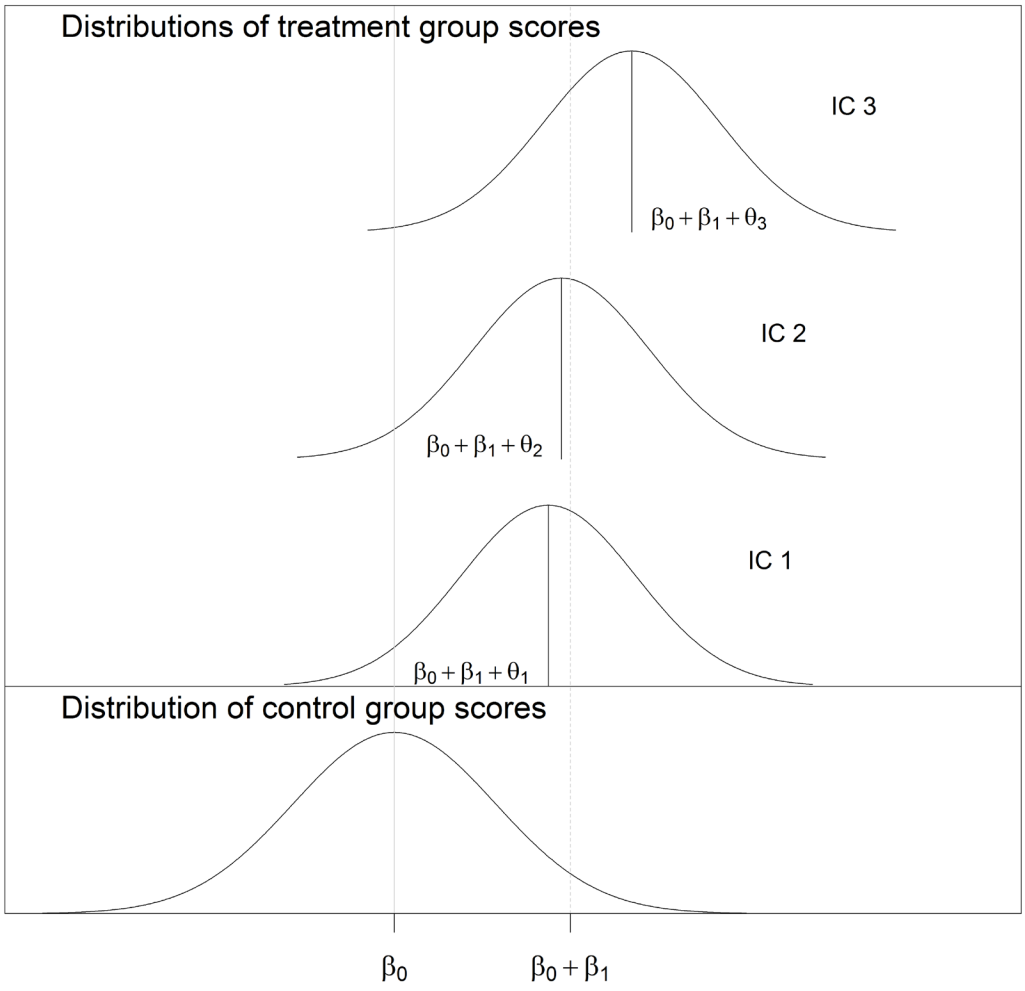
In equation (3.3), the IC effect $\theta_i$ occurs only in the treatment group, reflecting the partially nested structure. We assume that $\theta_i \sim N(0, \sigma^2_\theta)$ and that $\varepsilon_{ij} \sim N(0, \sigma^2_{\varepsilon T})$ for students in the treatment group ($i = 1$ to $I_T$) and $\varepsilon_{0j} \sim N(0, \sigma^2_{\varepsilon C})$ for students in the control group, and that all $\theta_i$ and $\varepsilon_{ij}$ are independent.[14]

Note that the model in equation (3.3) differs from an approach sometimes used in the literature where controls are treated as belonging to one "cluster" with effect $\theta_0 \sim N(0, \sigma^2_\theta)$. Bauer et al. (2008) explain why this approach (Approach 2 in their paper) is incorrect. It assumes that the control group students exhibit the same clustering as the treatment group students in an IC. The essential feature of a basic PN-RCT is that the treatment group has IC-level variance, while the control group does not. The model in equation (3.3) reflects that asymmetric structure by having the extra variability due to ICs in the treatment group only.

[14] Note that alternatively the model can be written as $y_{ij} = \beta_0 + \beta_1 T_{ij} + \theta_i + \varepsilon_{ij}$ if we have $\theta_i \sim N(0, \sigma^2_\theta)$ for ICs 1 to $I_T$ in the treatment group, and we prescribe $\theta_0 \sim N(0, 0)$ for the control group. Using a $N(0, 0)$ distribution forces $\theta_0$ to be equal to 0 and says that there is no additional variability due to clustering in the control group. This is the form of the model that we use when estimating model parameters using SAS software.

Figure 6 displays the distribution of student test scores from a basic PN-RCT, assuming arbitrarily that $\sigma_\theta^2 = 0.15\sigma_{\varepsilon T}^2$. Note that the control group scores all come from the same normal distribution but that the ICs in the treatment group have mean scores that deviate from the overall treatment group mean of $\beta_0 + \beta_1$. This causes students in the same IC to be more similar, on average, than two students from different ICs.

**Figure 6. Distribution of test scores for treatment and control groups for a PN-RCT.**




The total variability for a student's score differs for students in the treatment and control groups. In the control group, the total variability of a score is $\sigma_{\varepsilon C}^2$. In the treatment group, the total variability of a score is $\sigma_{\varepsilon T}^2 + \sigma_\theta^2$.

The generalized least squares estimator for $\beta_1$ in equation (3.3) and its variance are derived in appendix A. If ICs in the treatment group all have the same number of students (i.e., the design is *balanced*), this is the difference-in-means estimator $\hat{\beta}_1 = \bar{y}_T - \bar{y}_C$, where $\bar{y}_T$ is the mean score for all students in the treatment group and $\bar{y}_C$ is the mean score for all students in the control group. The variance of $\hat{\beta}_1$ is the sum of the variance for the treatment group mean when IC sizes are equal[15] $(\sigma_\theta^2 / I_T + \sigma_{\varepsilon T}^2 / n_T)$ and the variance for the control group mean $(\sigma_{\varepsilon C}^2 / n_C)$:

$$Var_{PN-RCT}(\hat{\beta}_1) = \frac{\sigma_\theta^2}{I_T} + \frac{\sigma_{\varepsilon T}^2}{n_T} + \frac{\sigma_{\varepsilon C}^2}{n_C}.$$

The variance of the treatment group will typically be larger than that for the control group because the IC variance component $\sigma_\theta^2 / I_T$ enters the variance expression for the treatment group only. This additional variability reflects the clustering from the ICs. There is an intraclass correlation only within the treatment group:

$$\rho_\theta = \frac{\sigma_\theta^2}{\sigma_{\varepsilon T}^2 + \sigma_\theta^2} \tag{3.4}$$

which represents the fraction of the total treatment group variation that is due to the variation in mean scores between ICs.

## 3.3 Constructed Data Example for the Basic PN-RCT Design

To illustrate how SAS software can be used to estimate the models from above, we consider an after-school pull-out reading intervention where treatment students are provided services in small groups. For our constructed example, we assume that there are 125 students in the control group and 125 students in the treatment group. We assume further that students in the treatment group are assigned to 25 ICs, each containing 5 students. Thus, we are considering a relatively small PN-RCT design with intensive, small-group instruction.

[15] For unequal IC sizes, the variance expression is given in equation (A.6) of appendix A.

We generated test score data using the random IC effects model in equation (3.3) under the assumption that the IC-level intraclass correlation coefficient, $\rho_\theta$, is 0.1. This intraclass correlation reflects differences in mean test scores across the ICs as a result of potential differences in the characteristics of students and teachers assigned to the ICs and the heterogeneity of treatment effects across ICs. The control group mean is assumed to be 100, and the treatment group mean is assumed to be 106, so the ATE in this example is 6 scale points. The variance due to students is assumed to be $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T} = 225 = 15^2$ for each group. Because $\sigma^2_\theta = \rho_\theta \sigma^2_{\varepsilon T} / (1 - \rho_\theta)$, these assumptions imply that $\sigma^2_\theta = 25$. The sample sizes of 125 students in each group were chosen so as to give 80 percent power to detect an impact of about 6 scale points, corresponding to an effect size of 0.4 standard deviations.[16] Appendix E gives the SAS code used to generate this data set.

The 250 observations for the students were graphed using SAS code displayed in figure 7. Table 3 gives the variable names and values for the data. The data are displayed graphically in figures 8 and 9. The simple side-by-side boxplot in figure 8 gives a rough idea of the relative means, medians, and variability of test scores for the treatment and control groups. As expected, the variability of test scores is greater for the treatment group because of the assumed IC variability. Figure 9 displays a more refined boxplot for the individual ICs in the treatment group that can be used to assess the variability of test scores across ICs and to detect ICs that have unusual values. As shown in figure 9, there are considerable differences in mean test scores across the ICs. Some of that variability is due to the residual variability among students, but some is also due to the intrinsic variability among the ICs.

**Table 3. Description of variables used in example**

| Name | Description | Type | Values |
|---|---|---|---|
| Y | Response variable | Numeric | |
| Trt | Indicator variable for treatment group (this is the $T_{ij}$ variable in equation (3.3). | Numeric | 0 for control group,<br>1 for treatment group |
| trtname | Character variable for treatment group, used in plots | Character | "Control" for control group<br>"Treatment" for treatment group |
| Ic | Intervention cluster | Numeric | 0 for control group,<br>1 – 25 for ICs in treatment group |
| Subjid | ID for each student, unique within each IC | Numeric | |

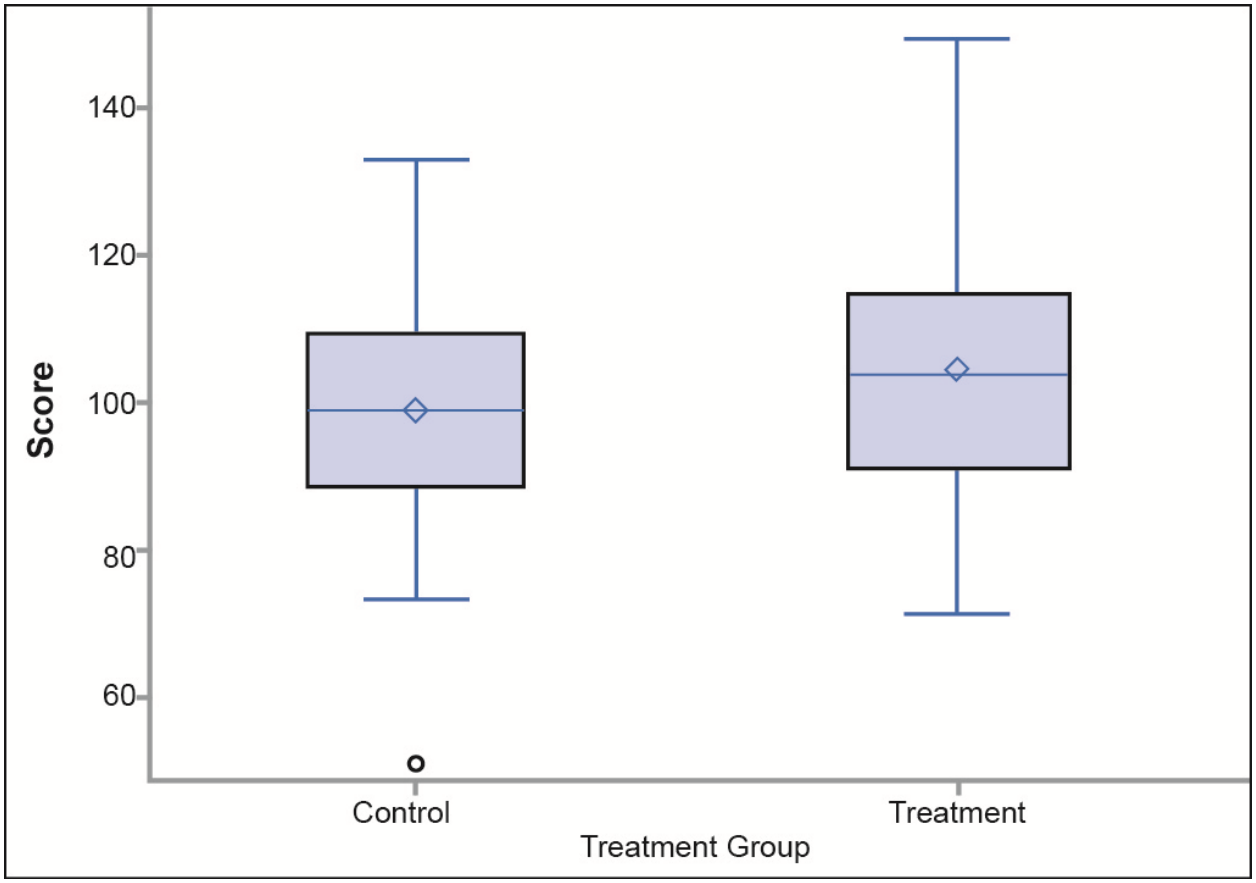
[16] Power calculations for this example are given in table 2; the sample size from that table for precision target 0.40, $\rho_\theta = 0.1$, average IC sample size = 5, and $R^2 = 0$ is 251.

Figure 7. SAS code to produce figures 8 and 9

```
proc sgplot data = model1;  /* Figure 8 */
   vbox y / category = trtname;
   yaxis label = "Score";
   xaxis label = "Treatment Group";
run;
/* Next plot assumes data are sorted by ascending value of IC median */
proc sgplot data = model1 noautolegend;  /* Figure 9 */
   vbox y / category=trtname group=ic grouporder=data;
   yaxis label = "Score";
   xaxis label = "Treatment Group";
run;
```

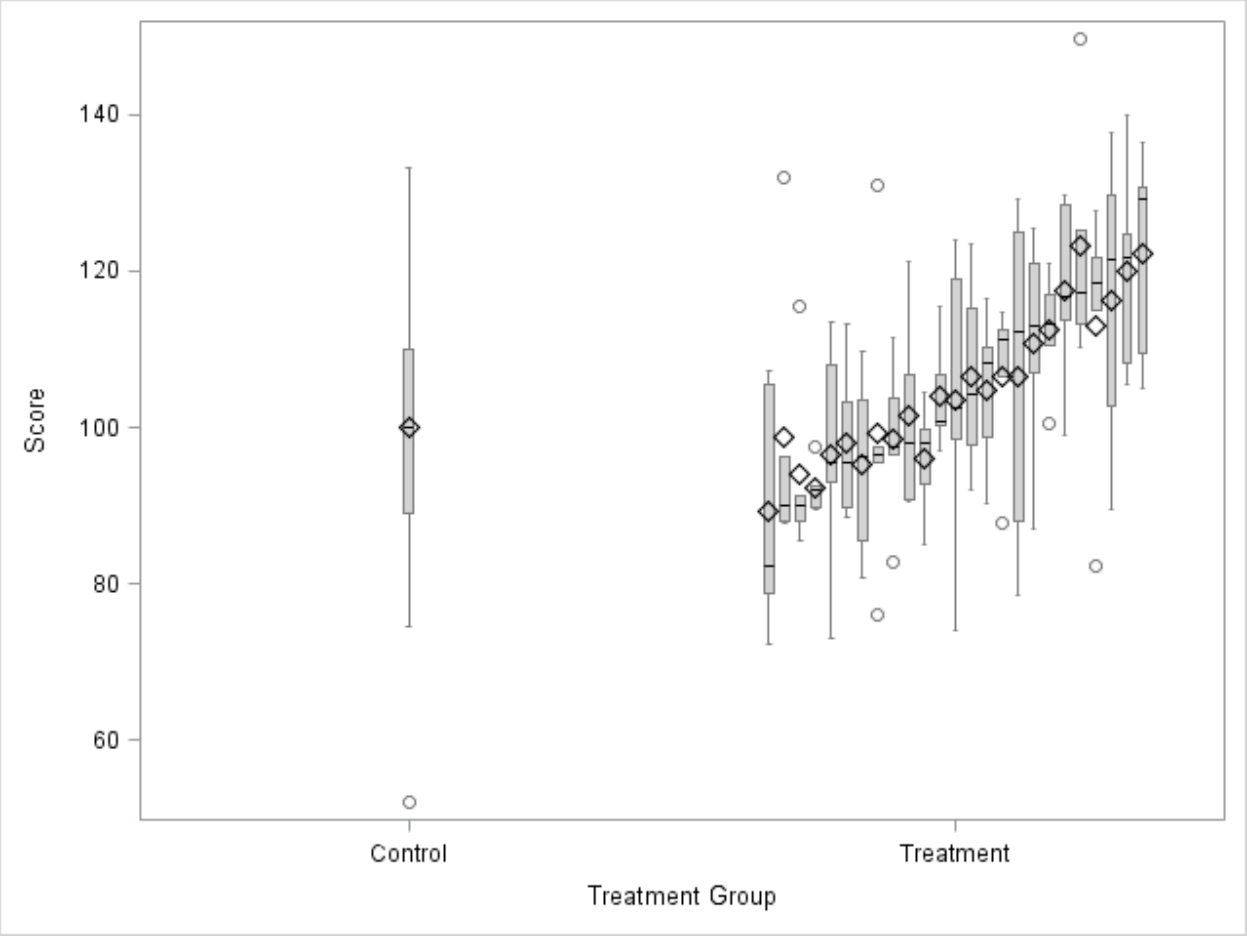
**NOTE:** The SAS code in figure 7 produces the basic plots. The full set of graphical options used to produce the shading and coloring is given in appendix E.

Figure 8. Boxplots of scores from control and treatment groups




**NOTE:** The median for each group is indicated by the center line in the box, and the mean is indicated by the diamond symbol. The first and third quartiles are the edges of the boxes (the length of the shaded box is the inter-quartile range). The lines extending above and below the boxes are the whiskers: the upper whisker extends from the third quartile to the maximum data point that is less than the third quartile + 1.5 x (inter-quartile range), and the lower whisker is defined analogously. The circles denote points whose distance below the first quartile or above the third quartile exceeds 1.5 x (inter-quartile range).

Figure 9. Plot of data, showing 1 boxplot for the control group and individual boxplots for the 25 ICs in the treatment group. The boxplots for the ICs in the treatment group are ordered by increasing value of the IC median




**NOTE:** The symbols used in this graph are described in figure 8.

After displaying the data, we are now ready to perform a statistical analysis. SAS PROC MIXED will analyze data from PN-RCTs by forcing the IC-level variability in the control group to equal 0. This is done through the PARMS statement, which is used in PROC MIXED to specify initial values for the covariance parameters. When the PARMS statement is used without options, PROC MIXED iterates, starting with the values for the covariance parameters given in the PARMS statement, to obtain final estimates for the covariance parameters. For a PN-RCT though, the IC-level variance in the control group is fixed at 0. This is done in PROC MIXED by specifying the initial value to be 0 and then placing a "hold" on that value, so it also forms the final estimate.

With large data sets, PROC MIXED can be slow to converge. The convergence can be speeded by using good initial estimates; this also helps ensure that the procedure converges to the correct values. You can obtain initial estimates of the residual variability by fitting a simple regression model to the

data, ignoring the IC effects (using PROC REG or PROC GLM); alternatively, for this design, you can obtain initial estimates by finding the variance of the control students and separately fitting a completely nested model to the treatment students. By doing this, we obtain initial estimates for $\sigma^2_{\varepsilon C}$, $\sigma^2_{\varepsilon T}$, and $\sigma^2_{\theta}$ of 197, 204, and 52 respectively. If no initial estimates are available, the non-zero variance components could be set to 1 as a default, but this will slow convergence.

The SAS code in figure 10 will give estimates of the treatment effect and the covariance parameters for Design 1 with random IC effects.[17] The code is general and can be modified to include baseline covariates and other model characteristics. It can be adapted to experiments with multiple treatment conditions by declaring "trt" to be a categorical variable in the CLASS statement.

The code exploits several features of SAS that would not typically be used when conducting a traditional C-RCT analysis. The first, as mentioned above, is the PARMS statement. The first variance component is fixed at 0; the others are estimated iteratively using initial values 52, 197, and 204, respectively. The ordering of the values in the PARMS statement is important: it follows the ordering of the variance components expressed in the RANDOM and REPEATED statements. The second feature is the use of the GROUP option, which tells SAS to fit separate variance components for each treatment group. The REPEATED statement calculates different student-level variances in the treatment and control groups. If you want to require $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T}$, then omit the REPEATED statement.

[17] Note that we use two separate variables for the treatment group: "trt" is a numeric variable taking on values 0 and 1, and "trtname" is a class variable, used to specify the groupings for the variance components. This is done so that the main model can be fit as a regression with independent variable "trt", and the estimate of $\beta_1$ is the ATE. The same p-values would result, however, if "trt" were replaced with "trtname" throughout: the only difference is that then the solution given by SAS is that for the estimate of $-\beta_1$ rather than $\beta_1$. The code in figure 10 provides the ATE, its standard error, and estimates of the variance parameters. It will work for both balanced and unbalanced data.

Figure 10. SAS code to fit mixed model for basic PN-RCT design

```
proc mixed data=model1;
  class subjid trtname ic;
  model y=trt / solution ddfm=sat;
     /* Fit a model in which the slope is the ATE */
  random intercept / group=trtname subject=ic(trtname);
     /* The random statement fits the hierarchical model in the treatment
        group; using the option group=trtname allows different variances in
        the control and treatment groups */
  parms (0) (52) (197) (204) / hold = 1;
     /* The parms statement gives initial values for the variance parameters;
        setting the first parameter to 0 with the hold statement ensures that
        no IC-level clustering is fit in the control group */
  repeated subjid/ group=trtname;
     /* The repeated statement allows different student-level variances to be
        fit in the control and treatment groups. This statement can be
        deleted if the student-level variances are assumed to be equal. */
  title 'Random Effects Estimator for Basic PN-RCT Model';
run;
```

When running the SAS code in figure 10, researchers will get the following message in the SAS log:

```
NOTE: Estimated G matrix is not positive definite.
NOTE: Asymptotic variance matrix of covariance parameter
estimates has been found to be singular and a generalized
inverse was used. Covariance parameters with zero variance
do not contribute to degrees of freedom computed by
DDFM=SATTERTH.
```

This is *not* an error message. SAS gives the note that the **G** matrix is not positive definite whenever a covariance parameter estimate is 0. This *always* occurs in PN-RCTs because the IC effect is deliberately set to 0 for the control group.[18] Do not worry if you see the above messages in your SAS log. You should worry, though, if you see a message that says:

```
WARNING: Did not converge
```

In that case, you may have misspecified the model or given poor initial values for the covariance parameters. The SAS log message

```
NOTE: At least one element of the gradient is greater
than 1e-3
```

[18] See appendix A for further discussion of the **G** matrix and positive definiteness in this model.

may also indicate a convergence problem. Kiernan, Tao, and Gibbs (2012) give useful guidance for diagnosing problems and error messages in PROC MIXED.

The following output is produced by the SAS code in figure 10. SAS first repeats the information you provided to it.

| Model Information | |
|---|---|
| Data Set | WORK.MODEL1 |
| Dependent Variable | Y |
| Covariance Structure | Variance Components |
| Subject Effect | ic(trtname) |
| Group Effects | trtname, trtname |
| Estimation Method | REML |
| Residual Variance Method | None |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Satterthwaite |

| Class level information | | |
|---|---|---|
| Class | Levels | Values |
| Subjid | 250 | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 |
| Trtname | 2 | Control Treatment |
| Ic | 26 | 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 |

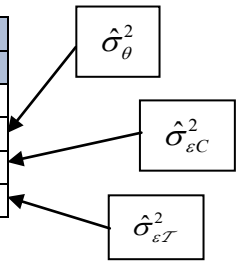
The code did not specify a method for estimating the covariance parameters, so SAS uses the default method of restricted maximum likelihood (REML). The option "method = ML" could be used if maximum likelihood estimators are preferred or if likelihood ratio tests are to be done comparing different models (see below).

SAS then gives you its understanding of the number of covariance parameters and other terms in the model. Here, there are four covariance parameters: 0 (for the IC effect in the control group), $\sigma^2_\theta$, $\sigma^2_{\varepsilon C}$, and $\sigma^2_{\varepsilon T}$. The subjects for this analysis are the ICs: there are 25 treatment ICs (coded 1 to 25) and one IC for the control group (coded as 0).

| Dimensions | |
|---|---|
| Covariance Parameters | 4 |
| Columns in X | 2 |
| Columns in Z Per Subject | 2 |
| Subjects | 26 |
| Max Obs Per Subject | 125 |

SAS then shows details of the iterative model fitting process, followed (if all is well) by the words "Convergence criteria met." The estimates of the covariance parameters are shown next.

| Covariance Parameter Estimates | | | |
|---|---|---|---|
| Cov Parm | Subject | Group | Estimate |
| Intercept | ic(trtname) | trtname Control | 0 |
| Intercept | ic(trtname) | trtname Treatment | 52.4786 |
| Subjid | | trtname Control | 197.13 |
| Subjid | | trtname Treatment | 204.32 |




As requested in the code, the IC variance in the control group is set to 0. The residual variance component in the control group, $\sigma^2_{\varepsilon C}$, is estimated to be 197.1. The IC-level variability and residual variability in the treatment group are estimated as 52.5 and 204.3, respectively. The $\hat{\sigma}^2_\theta$ value of 52.5 is larger by chance than the value of 25 used to construct the data because of sampling error due to relatively small numbers of ICs. The model in equation (3.3) captures the key features of the PN-RCT design. Most importantly, it incorporates the different variance structures in the treatment and control groups, which is crucial for obtaining the appropriate standard error of the ATE.

The final parts of the output are the solutions for the fixed effects. The intercept value of 99.9 is the test score mean for the control group students, and the estimated treatment effect is 5.16 scale points with a *p*-value of 0.0299. Thus, in this example, the reading intervention improved English language learner student outcomes by 5.16 scale points (which translates into an effect size of about 0.35 standard deviations), and this estimated ATE is statistically significant. Note that 46.9 degrees of freedom are used for this estimate. These are the degrees of freedom calculated using the Satterthwaite (1946) approximation, which is recommended when the two groups being compared have *unequal* variances. Details about the degrees of freedom (and why they can take on non-integer values) are given in appendix B.

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard error | DF | t Value | Pr > \|t\| |
| Intercept | 99.9174 | 1.2558 | 124 | 79.56 | <.0001 |
| Trt | 5.1610 | 2.3045 | 46.9 | 2.24 | 0.0299 |

ATE

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| trt | 1 | 46.9 | 5.02 | 0.0299 |

*p*-value for ATE

An analysis of the residuals and regression diagnostic statistics from this model (using the RESIDUAL and INFLUENCE options to the MODEL statement in PROC MIXED) revealed no unusual features except for one potentially outlying student in the control group (shown as the circle at the bottom of the control group boxplot in figures 8 and 9). Guides to regression diagnostics for mixed models are found in Schabenberger (2004), Jiang (2007, section 2.4), and SAS Institute (2011, section on Residuals and Influence Diagnostics).

Note that although this particular balanced data set may be analyzed more simply (as discussed in the next subsection), PROC MIXED and the structure given in figure 10 are needed for unbalanced data, when there are baseline covariates, or whenever information about the sources of variability is desired. A likelihood ratio test may be used to test certain hypotheses about the covariance structure. For this example, a hypothesis of interest is whether the residual variances are equal in the control and treatment groups ($H_0 : \sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T}$). To perform a likelihood ratio test, refit the full, unrestricted model using the code in figure 10 with maximum likelihood rather than the default restricted maximum likelihood method (add METHOD = ML to the PROC MIXED line). The value of-2 Log Likelihood from SAS is 2053.5. Then fit the restricted model where $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T}$ (by deleting the REPEATED statement in figure 10) with maximum likelihood. SAS gives the value 2053.6 for-2 Log Likelihood. The likelihood ratio test statistic is then 2053.6 – 2053.5 = 0.1. This value is compared to a chi-squared distribution with degrees of freedom equal to the number of restrictions (1 in our case), resulting in a *p*-value that is greater than 0.10, indicating that there is no significant difference between the two variance parameters. This is to be expected because we constructed our data assuming equal residual variances for the treatment and control groups, but data sets from other experiments may exhibit unequal variances. A similar test statistic can be used for all models that we consider below.[19]

[19] Note that SAS PROC MIXED also provides an option COVTEST for testing whether covariance parameters are equal to 0. This test, however, is based on a normal approximation to the distribution of the estimated variance component and, as stated in SAS Institute (2011), is inaccurate with small numbers of groups (see also Lohr and Divan 1997). Self and Liang (1987) and Skrondal and Rabe-Hesketh (2004) discuss methods that can be used to test whether variance components are 0. SAS PROC GLIMMIX (Version 9.2 and later) implements some of these methods.

## Alternative Methods for Estimating ATEs

The SAS code in figure 10 is general and can be modified to include baseline covariates and other model characteristics. This code, however, makes use of features unique to SAS software such as the PARMS statement. In this section, we present two methods that can be used with SAS or other software packages to estimate the ATE for the model given in equation (3.3). Both methods would need modification to work if the model has additional covariates.

Both methods exploit the fact that the ATE is $\hat{\beta}_1 = \bar{y}_T - \bar{y}_C$, where $\bar{y}_C$ is the mean test score of all students in the control group, and where $\bar{y}_T = n_T^{-1}\sum_{i=1}^{I_T} J_i \bar{y}_{Ti}$ is a weighted average of the individual IC means in the treatment group, weighted by the number of students in each IC.

The first method involves estimating separate models for the treatment and control groups and aggregating the findings. This can be performed by noting that in the basic PN-RCT design, $Var(\bar{y}_T - \bar{y}_C) = Var(\bar{y}_T) + Var(\bar{y}_C)$ because $\bar{y}_T$ and $\bar{y}_C$ are independent. We can estimate $Var(\bar{y}_C)$ by $\hat{\sigma}^2_{\varepsilon C} / n_C$ and can estimate $Var(\bar{y}_T)$ by fitting the one-way random effects model

$$y_{ij} = \mu_T + \theta_i + \varepsilon_{ij}$$

to the observations in the treatment group. This was, in fact, done earlier to obtain initial parameter estimates for the full SAS code in figure 10. For our example, $\widehat{Var}(\bar{y}_C)$ = 197.13/125 = 1.577 and $\widehat{Var}(\bar{y}_T)$ =3.734. Then, $\widehat{Var}\left(\bar{y}_T - \bar{y}_C\right) = 3.734 + 1.577 = 5.311$ and the test statistic for testing $H_0 : \beta_1 = 0$ is

$$T = \frac{\bar{y}_T - \bar{y}_C}{\sqrt{\widehat{Var}\left(\bar{y}_T - \bar{y}_C\right)}} = \frac{105.08 - 99.92}{\sqrt{5.311}} = 2.24.$$

This test statistic can be compared to a normal distribution, or, alternatively, the Satterthwaite degrees of freedom approximation to the *t*-test (appendix B) may be applied. Note that the test statistic is the same value as was produced by the code in figure 10. It just requires some extra calculation to assemble the component statistics.

The second method analyzes the data using IC *means* for the treatment group and the full data for the control group. This method gives an exact analysis for balanced data and an approximate

analysis for unbalanced data. This method relies on the model assumption that the IC means for the treatment group are independent observations. When each IC has the same number of students, $Var(\bar{y}_T) = Var(\bar{y}_{T1}) / I_T$ . This implies that we can analyze the data using a two-sample *t* test with $n_C$ independent observations in the control group (each with variance $\sigma^2_{\varepsilon C}$) and $I_T$ independent observations $(\bar{y}_{T1},\ldots,\bar{y}_{TI_T})$ in the treatment group (each with variance $\sigma^2_\theta + \sigma^2_{\varepsilon T} / J$). If you use this method for the analysis, make sure you do the *t* test with *unequal* (*not pooled*) variances, so you capture the extra variability in the treatment group arising from the ICs. The SAS code to perform this *t*-test is given in appendix E; the estimated ATE and its *p*-value are the same as in the full analysis produced by figure 10 for this example. This method can be used to provide an approximate analysis when IC sizes are unequal by assigning weight 1 to each control group observation and weight $J_i$ to IC group $i$.

Both methods depend on the identity $Var(\bar{y}_T - \bar{y}_C) = Var(\bar{y}_T) + Var(\bar{y}_C)$. When baseline covariates are included, the covariate-adjusted estimated treatment means from the two groups may be correlated, so these methods would not extend to that case.

## 3.4 Including Covariates in the Models

Baseline covariates are often used to analyze RCT data to improve the precision of the estimates and to adjust for residual treatment-control differences in baseline characteristics due to random sampling. Baseline covariates can be measured at the student, educator, school, or site level.

The basic PN-RCT analysis is easily modified to include other covariates. In SAS PROC MIXED, simply include the additional covariates in the MODEL statement. Thus, for example, to fit a model with student-level indicator variables for sex and free or reduced-price lunch status (labeled as frl), the MODEL statement from each design would be modified to read

```
model y=trt sex frl/solution ddfm=sat;
```

SAS will then automatically adjust the treatment effect and estimated variance components to account for the covariates. Interactions between covariates and the treatment effect may be of interest as well, and these may be fit by including the interaction terms `trt*sex` and `trt*frl` in the MODEL statement.

In experiments in which a pretest and a posttest are measured for all students, interest centers on whether the growth in the treatment group differs from the growth in the control group. This can

be evaluated in two ways: (1) by defining the response variable to be the gain score, with $y = posttest - pretest$, or (2) by defining the response variable to be the posttest score and including the pretest score as a covariate in the model (see Brogan and Kutner 1980; Dugard and Todman 1995; and Laird 1983 for discussion of the relative merits of the two approaches). The latter approach can be implemented by using the statement

```
model posttest=trt pretest/solution ddfm=sat;
```

The inclusion of model covariates will increase the precision of the ATE estimates because the random error terms are now conditional on the covariates. Because of random assignment, the covariates are uncorrelated with treatment status in expectation. Thus, in the presence of covariates, the variance components in the above variance expressions can be deflated by the factor $(1-R_q^2)$, where $R_q^2$ is the proportion of the total variance of the outcome at hierarchical level $q$ that is explained by the covariates. For instance, for the basic PN-RCT design with student-level randomization and random IC effects, the variance expression with covariates is

$$Var\left(Impact\right)=\left[\frac{\sigma_{\varepsilon T}^2\left(1-R_{\varepsilon T}^2\right)}{n_T}+\frac{\sigma_{\theta}^2\left(1-R_{\theta T}^2\right)}{I_T}+\frac{\sigma_{\varepsilon C}^2\left(1-R_{\varepsilon C}^2\right)}{n_C}\right],$$

where $R_q^2$ values can differ across three variance components.

The section on forming ICs in chapter 2 discussed that in some experiments, students may be tracked into ICs based on their pre-intervention characteristics. This sort of deliberate grouping can inflate the IC-level variability because the differences among ICs would be partly due to the differences in those student characteristics. Controlling for these preexisting differences using baseline covariates will help mitigate this problem and yield intraclass correlation estimates that reflect more policy-relevant variation across ICs due to differences in teacher quality, curriculum, and so on.

A baseline covariate that applies only to the treatment group (e.g., a covariate that describes a characteristic of the IC instruction) needs to be constructed carefully in SAS, so the covariate applies only to the treatment group and not to the control group. As an example, consider a situation for the basic PN-RCT design in which some of the ICs have experienced tutors and others have novice tutors. Fitting equation (3.3) with IC as a random factor will include the differences due to tutor experience (a fixed factor) in the IC-to-IC variability. Including a measure of IC tutor experience,

*IC_Exp*, as a covariate (where the value of the covariate for each student is the measure of experience associated with his or her IC tutor) will increase precision because the IC-level variance component will be the variability remaining after accounting for the effects of tutor experience.

We have to be careful about constructing the variable *IC_Exp* because it only pertains to students in the treatment group. One approach is to construct a centered *IC_Exp* variable, *C_IC_Exp*, that equals 0 for the control group and equals *IC_Exp* minus the mean of *IC_Exp* for the treatment group. This approach allows the variable *trt* to estimate the overall treatment effect because *C_IC_Exp* will be orthogonal to *trt*. More specifically, *C_IC_Exp* can be constructed as follows:

$$C\_IC\_Exp_{ij} = \begin{cases} 0 & \text{if student } j \text{ is in the control group } (i = 0) \\ IC\_Exp_{ij} - \overline{IC\_Exp} & \text{if student } j \text{ in IC } i \text{ is in the treatment group} \end{cases}$$

where $\overline{IC\_Exp}$ is the mean value of the IC tutor experience measure across all students in the entire treatment group. The following SAS code will estimate parameters in this model:

```
proc mixed data=model1 ;
  class subjid trtname ic;
  model y=trt c_ic_exp / solution ddfm=sat ;
  random intercept / group=trtname subject=ic(trtname);
  parms (0) (50) (200) / hold = 1;
  title 'IC random effects nested in fixed factor';
```

When including IC-level covariates, it is important to have sufficient ICs in the experiment to allow assessing the significance of those covariates. In section 1.2.1, we stated that a PN-RCT must have at least two ICs to be able to estimate the IC-level variability. If there are $k$ IC-level covariates, the experiment must have at least $k+2$ ICs to be able to assess their significance.

Finally, it is important to recognize that valid model covariates must be measured at baseline and not have been affected by the intervention. This issue is complex for IC-related covariates because ICs are typically formed after random assignment. When students are randomly assigned to ICs, then including IC-level covariates such as tutor experience accounts for some of the IC-to-IC variability. When students are purposively assigned to ICs, however, caution is needed. If, for example, the students viewed as most in need of help are assigned to the most experienced tutors, a covariate for tutor experience would also be including effects of a baseline characteristic of the students in the IC. The tutor experience covariate would then be partially controlling for the student characteristic in the treatment group but there would be no comparable covariate for the control group.

## 3.5 The Blocked PN-RCT Design

In many PN-RCTs, randomization will be done separately within sites or schools. The BELL evaluation in example 1.4 was of this type. Thus, for the BELL study, students in Boston were randomized to treatment or control, and students in Chicago were randomized to treatment or control. The basic PN-RCT design was, thus, carried out independently in each site. The control students in each site are from the same school system as the treatment students. Because there are multiple sites, the results can be generalized to the types of sites that participated in the experiment.

The difference between the basic PN-RCT and a blocked PN-RCT hinges on *where the randomization is performed*. In the basic PN-RCT, students in the full population are randomly assigned to the treatment or control condition. In this design, it is possible for *any number* of students originating from a specific school or neighborhood to end up in the treatment group. If, however, schools are used as a blocking unit, the randomization is done separately for each school. In this design, exactly half the students in each school would be randomly assigned to the treatment group, and the other half would be randomly assigned to the control group.

Note that students within the same block are generally positively correlated because they share the same environmental factors. Sometimes this can generate confusion between blocks and clusters; in some experiments, schools might serve as blocks, while in other experiments, schools might be clusters. Here is the difference: in a blocked design, the positive correlation generally increases the precision of the ATE because half of the students in each block receive the treatment and the other half receive the control. Comparing the treatment and control students within a block subtracts out the effects of common environmental factors from being in the same school. In a C-RCT with schools as clusters, *all* students in the school receive either the treatment or control, so the positive correlation decreases the precision of the ATE.

One consideration when contemplating a blocked PN-RCT is whether there are likely to be "spillover" or contamination effects between the treatment and control students within a block. Suppose that students are randomly assigned to treatment and control groups within classes: the treatment students in each class are pulled out into ICs after school to learn about better study habits, while the control students in the class receive no additional instruction. The students in the control group will know that the treatment students are getting a special intervention, and they will very likely learn something about that intervention. Several possible contamination scenarios can be

imagined. The control students might learn about the helpful study habits from their friends in the treatment group and start adopting them, thereby reducing the observed treatment effect. There might be a negative contamination effect in other experiments. Students might not understand that the treatment students were chosen randomly for the intervention and may think there is a specific reason that some students were chosen while others were not. It is possible that students in the control group (not chosen for the intervention) may be demoralized and, therefore, perform worse than they would in a classroom not selected for the study; this would produce an estimated treatment effect that is exaggerated.

Potential contamination should be considered when deciding whether to use a blocked PN-RCT (Jenney and Lohr, 2009; Moerbeek 2005). In general, contamination is less likely if the blocking units are larger (e.g., districts, neighborhoods, or cities) than if they are small (e.g., classrooms). If there is concern that contamination effects can be large, it might be better to use a cluster-randomized design instead of a blocked PN-RCT, as discussed in section 4.1.

### 3.5.1 Model and Model Implications

To help make the concepts concrete, assume that the blocking factor is a school. A total of $H$ schools are available for the experiment, and students are randomized to control or treatment group separately within each school. Following the initial randomization to treatment or control, the students in the treatment group are then assigned to one of the ICs formed within the school.

We first consider models where school effects are treated as random and then consider models where school effects are treated as fixed. If school effects are treated as random, it is assumed that the study schools are randomly sampled from a broader population of schools, so study results generalize to this population. Under the fixed effects assumption, the study schools are assumed to be representative of themselves only. These two scenarios on handling block effects lead to considerably different estimation models.

**Treating School Effects as Random.** To help provide intuition on the estimation methods for the blocked PN-RCT design, we first describe an analysis for a *balanced* design, in which each school has the same number of treatment students and the same number of control students, and each IC has the same number of students. Let $h$ be the index for the school, let $\bar{y}_{Th}$ denote the mean score of the treatment group students in school $h$, and let $\bar{y}_{Ch}$ denote the mean of the control group students in school $h$. The ATE is given by

$$\hat{\beta}_1 = \overline{y}_T - \overline{y}_C = \frac{1}{H}\sum_{h=1}^{H}(\overline{y}_{Th} - \overline{y}_{Ch}) = \frac{1}{H}\sum_{h=1}^{H} u_h.$$

Note that $u_h = (\overline{y}_{Th} - \overline{y}_{Ch})$ is the estimated treatment effect in school $h$. We thus estimate the overall effect of the treatment by averaging the $H$ separate school treatment effects. In essence, we have $H$ independent replications of the basic PN-RCT design and average their results to calculate the ATE.

The schools are independent units, so in this balanced design the variance of the ATE may be estimated by

$$\widehat{Var}_{Blocked\ PN-RCT}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}_u^2}{H}$$

where $\hat{\sigma}_u^2 = \frac{1}{H-1}\sum_{h=1}^{H}(u_h - \overline{u})^2$ is the sample variance of the $H$ individual schools' estimated treatment effects. Note that the estimate $\hat{\sigma}_u^2$ will include the IC-level variability as well as the variability in treatment impact across schools. Essentially, we can evaluate the treatment impact in a balanced design by using a one-sample *t* test with $H-1$ degrees of freedom, where the "observations" are the observed treatment effects $u_h = (\overline{y}_{Th} - \overline{y}_{Ch})$ for each of the $H$ schools.

We need a more complicated setup if the experiment is not perfectly balanced, if it is desired to include covariates in the analysis model, or if there is interest in the estimates of the different variance components. The remainder of this section describes the general model for a blocked PN-RCT and may be skipped by readers not interested in the technical details.

Let the subscript $h$ represent the school, $i$ represent the IC within the school, and $j$ represent the student. The following notation is used:

$$y_{hij} = test\ score\ of\ student\ j\ in\ IC\ i\ of\ school\ h$$
$$\xi_h = effect\ of\ school\ h$$
$$\theta_{hi} = effect\ of\ IC\ i\ in\ school\ h$$
$$\varepsilon_{hij} = student-level\ error\ (residual)\ term\ for\ student\ j\ in\ IC\ i\ of\ school\ h.$$

The subscripts take values $h=1$ to $H$ for schools, $i=0$ to $I_{Th}$ for students in school $h$ (as for the basic PN-RCT, IC 0 is the value assigned to the control students, who are not in an IC, and ICs 1 through $I_{Th}$ are the intervention clusters for the treatment students within school $h$). The students

within the control group in each school are indexed by $j = 1$ to $J_{h0}$. The treatment group students in IC $i$ of school $h$ are indexed by $j = 1$ to $J_{hi}$. There are thus a total of $n_{Ch} = J_{h0}$ students in the control group in school $h$, and a total of $n_{Th} = \sum_{i=1}^{I_{Th}} J_{hi}$ students in the treatment group in school $h$.

The assignment to the treatment or control group differs for different students within the same school, so we let the treatment indicator depend on $h$, $i$, and $j$:

$$T_{hij} = 1 \text{ if student } j \text{ in IC } i \text{ of school } h \text{ is in the treatment group, } 0 \text{ if in the control group}$$

To help fix concepts, we first present a simplified model that assumes that the treatment impact is the same for all schools. In this case, the model for the blocked PN-RCT design is

$$y_{hij} = \beta_0 + \beta_1 T_{hij} + \xi_h + \theta_{hi} T_{hij} + \varepsilon_{hij}. \tag{3.5}$$

As before, we assume that $\varepsilon_{hij} \sim N(0, \sigma^2_{\varepsilon C})$ for students in the control group and that $\theta_{hi} \sim N(0, \sigma^2_{\theta})$ and $\varepsilon_{hij} \sim N(0, \sigma^2_{\varepsilon T})$ for ICs and students in the treatment group, respectively. The school effect pertains to students in *both* the treatment and control groups, so we assume $\xi_h \sim N(0, \sigma^2_{\xi})$ for $h = 1$ to $H$.

Let's now look at the estimated treatment effect for school $h$, which is $\bar{y}_{Th} - \bar{y}_{Ch}$ under the model in equation (3.5). We have

$$\begin{aligned}
\bar{y}_{Th} - \bar{y}_{Ch} &= \frac{1}{n_{Th}} \sum_{i=1}^{I_{Th}} \sum_{j=1}^{J_{hi}} y_{hij} - \frac{1}{n_{Ch}} \sum_{j=1}^{n_{Ch}} y_{h0j} \\
&= \frac{1}{n_{Th}} \sum_{i=1}^{I_{Th}} \sum_{j=1}^{J_{hi}} \left( \beta_0 + \beta_1 T_{hij} + \xi_h + \theta_{hi} T_{hij} + \varepsilon_{hij} \right) - \frac{1}{n_{Ch}} \sum_{j=1}^{n_{Ch}} \left( \beta_0 + \xi_h + \varepsilon_{h0j} \right) \\
&= \beta_1 + \frac{1}{n_{Th}} \sum_{i=1}^{I_{Th}} \sum_{j=1}^{J_{hi}} \left( \theta_{hi} T_{hij} + \varepsilon_{hij} \right) - \frac{1}{n_{Ch}} \sum_{j=1}^{n_{Ch}} \varepsilon_{h0j}.
\end{aligned}$$

Because each school has both treatment and control students, the school-level terms $\xi_h$ *cancel* for the individual school-level treatment effects. For balanced designs, then, the school-to-school variability will not appear in the variance of the estimated treatment effect.

A more general formulation of the model for this design (which is equivalent for balanced data to the paired *t* test analysis described above) allows treatment effects to vary across schools. In that

case, there may be an interaction effect between the school and the treatment (see Gates 1995; Lohr 1995; and McLean, Sanders, and Stroup 1991 for a discussion of interaction terms in mixed models). If these potential treatment-by-school interaction effects are assumed to be random, we obtain the following random coefficient regression model:

$$y_{hij} = \beta_0 + \beta_1 T_{hij} + \xi_h + \eta_h T_{hij} + \theta_{hi} T_{hij} + \varepsilon_{hij}. \qquad (3.6)$$

In equation (3.6), the random effect $\eta_h$ is an additional source of variability in school $h$ that could arise, for example, because of potential varying implementations or supports across schools. We consider the school-level effects $(\xi_h, \eta_h)$ to follow a bivariate normal distribution with $Var(\xi_h) = \sigma_\xi^2$, $Var(\eta_h) = \sigma_\eta^2$, and $Cov(\xi_h, \eta_h) = \sigma_{\xi\eta}$.

The ATE variance structure for this general design is derived in appendix A. The variance takes a special form in a balanced design where each school has the same number of treatment and control students, and each IC has the same number of students. In that case, letting $I_T$ be the number of ICs in each school, $n_T$ be the total number of treatment students (across all schools), and $n_C$ be the total number of control students, we find that

$$Var(ATE) = \frac{\sigma_\eta^2}{H} + \frac{\sigma_\theta^2}{H \times I_T} + \frac{\sigma_{\varepsilon T}^2}{n_T} + \frac{\sigma_{\varepsilon C}^2}{n_C}.$$

The leading variance term in this expression pertains to the extent to which treatment effects vary across schools.

**Treating School Effects as Fixed.** In some studies, it may be preferable to treat school effects as fixed effects. This would be the case if schools were chosen deliberately to meet some criterion: for example, school 1 is a large urban school, school 2 concentrates on arts education, and so on. Fixed school effects may also be preferred when there are only a few schools (or sites). In this case, it is difficult to think of the study schools as representative of a population of schools. In the basic PN-RCT, the school (or site) effect is always treated as fixed; variability among schools cannot be considered when there is only one school in the study, so the study school is considered to represent itself alone.

Models in which schools are fixed effects may be fit by constructing indicator variables for the school effects and their interactions with treatment, then running the regression model. In essence,

impact and variance estimates are obtained for *each* school and are then averaged to obtain overall estimates. In this model, the ICs may be treated either as fixed (the likely scenario given that school effects are being treated as fixed) or as random.

### 3.5.2 Constructed Data Example for the Blocked PN-RCT

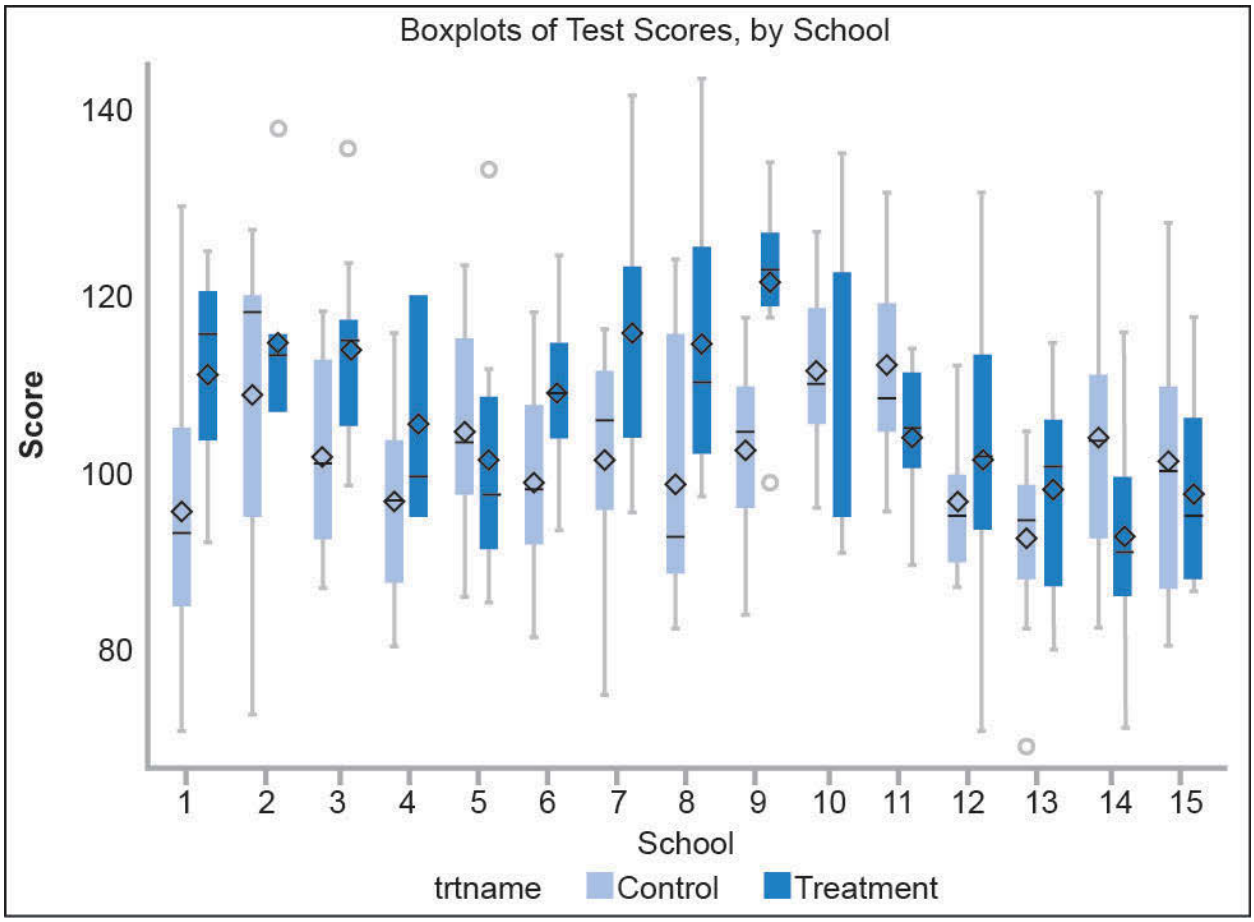
For this example, we assume that 15 schools participate in the study, and 20 students participate from each school. We assume that schools are random blocks. In every school, 10 of the students are randomly assigned to the control group; the other 10 are randomly assigned to 1 of 2 ICs (each with 5 students) in the school. The variance parameters are also set to the same values, with $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T} = 175$ for the student-level (residual) variance in each group, $\sigma^2_{\xi} = 45$, and $\sigma^2_{\theta} = 25$. For this example, we set $\sigma^2_{\eta} = 10$ and $\sigma_{\xi\eta} = 0$, so we can illustrate the features of the model in equation (3.6). The data were generated with a mean of 100 scale points for students in the control group and 106 scale points for students in the treatment group.

Figure 11 gives SAS code used to produce the plots of the data shown in figure 12. In this design, we want to examine the difference between treatment and control students separately within each school. Thus, we look at side-by-side boxplots separately for each school in figure 12. From figure 12, it can be seen that the treatment student mean is higher than the control student mean within 10 of the schools; in the other 5 schools, the control student mean is higher. There is also substantial variability among the schools; some schools have both treatment and control groups that are high, and other schools have both treatment and control groups that are low. However, as shown above, this school-to-school variability is removed from the estimated treatment effect in the blocked PN-RCT because it pertains to both treatment and control groups.

**Figure 11. SAS code used to construct figure 12**

```
proc sgplot data=model2;  /* Figure 12 */
   vbox y / category=school group=trtname meanattrs=(symbol=Diamond)
            medianattrs = (color=black);
   yaxis label= 'Score';
   xaxis label = 'School';
   title 'Boxplots of Test Scores, by School';
```

Figure 12. Boxplots of test scores for control and treatment students for each school




**NOTE:** The symbols used for the boxplots are defined in figure 8.

For this balanced design, the overall ATE for the experiment is the average of the 15 individual within-school ATEs, and its standard error is the sample standard deviation of those values divided by $\sqrt{15}$. A simple analysis, therefore, uses a one-sample *t* test on the values of the individual estimated treatment effects $(\bar{y}_{Th} - \bar{y}_{Ch})$ from the 15 schools. These 15 values are:

| School | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $(\bar{y}_{Th}-\bar{y}_{Ch})$ | 15.36 | 5.81 | 11.84 | 8.51 | -3.34 | 9.74 | 13.68 | 15.51 |

| School | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|
| $(\bar{y}_{Th}-\bar{y}_{Ch})$ | 18.49 | -1.52 | -7.94 | 4.90 | 5.43 | -11.11 | -3.48 |

The mean of the 15 values is ATE= 5.4608, and the standard deviation of the 15 values is 9.1228. The *t* statistic for the ATE is, therefore, $\sqrt{15}\,(5.4608)/9.1228 = 2.32$, which, when compared to a *t*

distribution with 14 degrees of freedom, gives a *p*-value of 0.0361. The output from SAS PROC TTEST is given in figure 13.

The SAS code in figure 14 is used to calculate the overall ATE and its standard error for the model in equation (3.6) and applies to unbalanced as well as balanced designs. The code provides estimates of variance components for the blocking (school) effects, the IC effect, and the control and treatment group residual variances. The initial parameter estimates are obtained by fitting a preliminary mixed model without the IC effects and with a common student-level variance.

**Figure 13. Output from SAS Proc TTEST, performing a *t* test on the 15 individual school ATEs.**

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 15 | 5.4608 | 9.1228 | 2.3555 | -11.1059 | 18.4901 |

| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| 5.4608 | 0.4087 | 10.5128 | 9.1228 | 6.6791 | 14.3876 |

| DF | t Value | Pr > \|t\| |
|---|---|---|
| 14 | 2.32 | 0.0361 |

**Figure 14. SAS code for estimating parameters in equation (3.6).**

```
proc mixed data=model2 noclprint;
   class trtname school ic subjid;
   model y = trt/ ddfm = sat solution cl;
   /* First random statement: Fit random coefficient regression model */
   random intercept trt/ subject=school type=un;
   /* Second random statement: Random effect of ICs,
      only for treatment students */
   random intercept/ group=trtname subject=ic(trtname school);
   /* Allow separate student-level variances */
   repeated subjid /group=trtname ;
   parms (15) (-10) (50) (0) (8) (160) (155)/ hold = 4;
   title 'Random Block PN-RCT';
```

We again suppress the output relating to class levels and iteration history (although you should check this in practice to make sure nothing is amiss). After fitting the model, we first examine the estimated variance components.

| Covariance Parameter Estimates | | | |
|---|---|---|---|
| **Cov Parm** | **Subject** | **Group** | **Estimate** |
| UN(1,1) | School | | 15.6042 |
| UN(2,1) | School | | -9.8197 |
| UN(2,2) | School | | 48.1723 |
| Intercept | ic(trtname*school) | trtname Control | 0 |
| Intercept | ic(trtname*school) | trtname Treatment | 5.2726 |
| Subjid | | trtname Control | 162.30 |
| Subjid | | trtname Treatment | 161.87 |

For this example, SAS estimates $\hat{\sigma}^2_{\varepsilon C} = 162.30$, $\hat{\sigma}^2_{\varepsilon T} = 161.87$, and $\hat{\sigma}^2_{\theta} = 5.27$ for the residual and IC-level variance components. The estimates for the school-level variance components are $\hat{\sigma}^2_{\xi} = 15.6$, $\hat{\sigma}^2_{\eta} = 48.1$, and $\hat{\sigma}^2_{\xi\eta} = -9.8$. The option TYPE = UN in the first RANDOM statement in figure 14 allows the school-level effects for the slope and intercept to be correlated. The values of the estimated variance components are imprecise because of the small number of schools.[20]

The estimated treatment effect for this example is 5.46 scale points, with *p*-value 0.0361. Note that the degrees of freedom used for the treatment effect is (number of schools) - 1. With the full model from equation (3.6), this analysis for balanced data has exactly the same ATE, *p*-value, and degrees of freedom as the *t* test using the 15 school-level ATEs as observations, given above.

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| **Effect** | **Estimate** | **Standard error** | **DF** | **t Value** | **Pr > \|t\|** |
| Intercept | 102.21 | 1.4568 | 14 | 70.16 | <.0001 |
| Trt | 5.4608 | 2.3555 | 14 | 2.32 | 0.0361 |

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| **Effect** | **Num DF** | **Den DF** | **F Value** | **Pr > F** |
| Trt | 1 | 14 | 5.37 | 0.0361 |

[20] You can also perform a formal hypothesis test for whether specific variance parameters are 0. See Self and Liang (1987) for methods that can be used to test hypotheses that are on the boundary of the parameter space in hierarchical models.

# Clustered PN-RCT Designs and Power Analyses 4

Chapter 3 described the statistical models for the basic and blocked PN-RCTs and showed how to analyze data sets for those models using SAS software. In this chapter, we consider other types of PN-RCT designs and provide detailed information on how to conduct power analyses for all the designs. Section 4.1 describes how to analyze data from the designs introduced in section 1.2.3, where naturally occurring clusters such as schools or classrooms are randomly assigned to the treatment or control condition, and ICs are formed within the clusters randomized to the treatment group. For these designs, both treatment and control groups have clustering, but the treatment group has an extra level of variability not found in the control group. Section 4.2 discusses cross-nested designs, where the ICs can cut across clusters or blocks. Finally, section 4.3 presents methods for calculating power for the different PN-RCT designs discussed in this paper.

## 4.1 Clustered RCTs With ICs Formed in the Treatment Group

Example 1.7 discussed the Number Rockets evaluation, in which 76 schools were randomly assigned to the treatment group or the control group. At-risk first-grade students in the treatment schools were assigned to different ICs that met after school; at-risk first-grade students in the control schools received no intervention. In this design, the schools are clusters, and the entire clusters are randomized to the treatment or control group. All participating students in a control school receive the control protocol, and all participating students in a treatment school receive the treatment protocol. This design, therefore, is a C-RCT because *clusters* are randomly assigned to the two study arms. However, this design can also be considered to be in the class of PN-RCT designs because the students were placed into small tutoring groups (the ICs). Thus, clustering effects differed for the treatment and control groups, because the treatment students received the intervention in a group setting, while the control students did not have that extra level of grouping; thus, this design has the distinguishing asymmetric design structure of PN-RCTs.

Note the difference between this design and the blocked PN-RCT design described in section 3.5. In the blocked PN-RCT design, half of the students in each school are in the treatment group, and the other half are in the control group. The shared environmental factors for that school therefore affect the treatment and control students in the school equally. Those environmental factors largely

cancel out when we look at the difference between the treatment and control students in that school. In the *clustered* design considered in this section, however, *all* the students in a school are in the same study arm (treatment or control). Therefore, the school environment affects all students in a specific school in the same direction: as discussed in section 3.1, this means that the variability between schools increases the variance of the ATE.

This clustered design, therefore, yields less precise estimates of the ATE than does a blocked PN-RCT. It has some major advantages for implementation, however. The contamination effects discussed in section 3.5 are not a concern for this design because all students in the school receive the same protocol, and this will help to minimize contact between treatment and control students. The clustered design may also have other advantages. For instance, it could increase administrative efficiency for implementing the intervention and could enhance study recruitment or student and teacher compliance because all students and educators in the school will share the same treatment status. Furthermore, this design might be more similar to how the intervention would be implemented on a widespread scale if found effective. These benefits of the clustered design could outweigh its reduced precision for estimating the ATE. Note that precision could be increased using regression models that include baseline covariates measured at the cluster level (the most important variables for increasing power) and at the student level.

When schools are randomized to treatment or control, there is clustering due to the schools in both arms of the study. The novel feature caused by the ICs is that the variability of a school mean is expected to be larger in a treatment school than in a control school because the treatment schools have an extra source of variability from the IC formation.

Although this section focuses on PN-RCT designs with school-level random assignment, similar methods apply to PN-RCT designs with classroom-level random assignment. An example of such a design is if classrooms within schools are randomly assigned to a treatment or control group and the intervention (e.g., a pull-out math program) is provided to students in treatment classrooms only. More generally, the methods we discuss in this section pertain to evaluations where education groups (such as schools, classrooms, districts, or neighborhoods) are the units of randomization.

### 4.1.1 Model and Model Implications

For the clustered design, let the subscript $h$ represent the school, $i$ represent the IC within the school (for schools randomized to the treatment group), and $j$ represent the student. We assume

that there are $H$ schools, with $H_C$ control schools ( $h = 1,\ldots,H_C$ ) and $H_T$ treatment schools $(h = H_C + 1,\ldots,H)$. We use the same convention as in section 3.2 to deal with the asymmetry of the two research groups arising from the partially nested structure, letting $i = 0$ refer to the "fictional" ICs in control group schools and $i$ =1 to $I_{Th}$ for the "real" ICs in school $h$ in the treatment group. For students, we have $j = 1$ to $J_{h0}$ for control group students in school $h$ and $j = 1$ to $J_{hi}$ for treatment group students in IC $i$ in treatment school $h$. The total number of control students is $n_C = \sum_{h=1}^{H_C} J_{h0}$ and the total number of treatment students is $n_T = \sum_{h=H_C+1}^{H} \sum_{i=1}^{I_{Th}} J_{hi}$. Let

$y_{hij}$ = test score of student *j* in IC *i* and school *h*
$T_h$ = 1 if school *h* is in the treatment group, 0 otherwise
$\theta_{hi}$ = effect of IC *i* in school *h*
$\varepsilon_{hij}$ = student-level error (residual) for student *j,* in IC *i* and school *h,*

where the treatment indicator $T_h$ depends only on the school (*h*) because randomization is done at the school level.

Combining the treatment and control group models yields the following unified model for the clustered design:

$$y_{hij} = \beta_0 + \beta_1 T_h + \xi_h + \theta_{hi} T_h + \varepsilon_{hij}. \qquad (4.1)$$

In this model, we assume that $\xi_h \sim N(0, \sigma^2_{\xi C})$ and $\varepsilon_{hij} \sim N(0, \sigma^2_{\varepsilon C})$ for schools and students in the control group, respectively, and that $\xi_h \sim N(0, \sigma^2_{\xi T})$, $\theta_{hi} \sim N(0, \sigma^2_{\theta})$, and $\varepsilon_{hij} \sim N(0, \sigma^2_{\varepsilon T})$ for schools, ICs, and students in the treatment group, respectively. As in the basic PN-RCT design, a special case of this model sets $\sigma^2_{\xi C} = \sigma^2_{\xi T}$ and $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T}$. We allow different variances for the school effects in the treatment and control groups because of potential heterogeneity of treatment effects across schools beyond those caused by IC-level variability.

The model in equation (4.1) has a very similar structure to the C-RCT model in equation (3.2). The cluster-level random effect in equation (4.1) is $\xi_h$, and the student-level error term is $\varepsilon_{hij}$. The only difference is that equation (4.1) has an extra term $\xi_h + \theta_{hi} T_h$ for the added variance in the treatment group caused by the IC-to-IC variability. This modification means that the total variability for a student's score differs for students in the treatment and control groups. In the control group, the

total variability of a score is $(\sigma^2_{\xi C} + \sigma^2_{\varepsilon C})$. In the treatment group, the total variability of a score is $(\sigma^2_{\xi T} + \sigma^2_{\varepsilon T} + \sigma^2_{\theta})$.

The model in equation (4.1) appears similar to the model in equation (3.5) for the blocked PN-RCT. However, there is a fundamental difference. In the cluster-randomized design, every student receiving the control protocol is independent of every student receiving the treatment protocol. In the blocked PN-RCT, however, treatment and control students in the same school are positively correlated because they share environmental factors specific to that school.

The variance expression for the mixed model ATE estimator for $\beta_1$ in equation (4.1) is derived in appendix A. For balanced designs, this variance is the sum of the variances for the treatment and control group means:

$$Var\left(Impact\right) = \left[\frac{\sigma^2_{\xi T}}{H_T} + \frac{\sigma^2_{\theta}}{H_T \times I_{TH}} + \frac{\sigma^2_{\varepsilon T}}{n_T}\right] + \left[\frac{\sigma^2_{\xi C}}{H_C} + \frac{\sigma^2_{\varepsilon C}}{n_C}\right].$$

Note that the school-level variance term enters both the treatment and control group variance expressions and will typically be the largest variance term because it is divided by the number of schools, which is likely to be considerably smaller than the total number of ICs or students. The extent to which these school-level terms will inflate the variances will depend on the intraclass correlation at the school level, which measures the extent to which mean scores vary across schools.

As discussed in section 3.1 for a C-RCT, school effects in the clustered design must *always* be treated as random; there is no fixed effects version of this model. Note that the observed variability across treatment school means will incorporate the additional variability caused by the presence of ICs. This result occurs because in multi-stage clustered designs, the *highest* level of clustering drives the variance estimates.

To demonstrate this result more formally for a balanced design, let $\overline{y}_{Th}$ be the mean score for treatment school $h$, and let $\overline{y}_T$ be the overall treatment group mean. Then, using equation (4.1), we find that the $\overline{y}_{Th}$ means, for $h = H_C + 1$ to $H$, are independent and identically normally distributed with mean $E[\overline{y}_T]$ and variance $Var\left(\overline{y}_{Th}\right) = [\sigma^2_{\xi T} + \frac{\sigma^2_{\theta}}{I_T} + \frac{\sigma^2_{\varepsilon T}}{I_T J}]$. An unbiased estimator for $Var\left(\overline{y}_{Th}\right)$ is the sample variance of treatment school means:

$$Tvar = \frac{1}{H_T - 1} \sum_{h=H_C+1}^{H} (\bar{y}_{Th} - \bar{y}_T)^2 \qquad (4.2)$$

and that an unbiased estimator for $Var(\bar{y}_T)$ is $(Tvar / H_T)$. Thus, in expectation, the variability among the school means in equation (4.2) will capture all sources of variation, *including* the IC-level variability and the student-level variability. Similarly, an unbiased estimator for $Var(\bar{y}_C)$ for the control group is the sample variance of the control school means, $Cvar$, divided by $H_C$. Because the schools in the control and treatment groups are independent, $Var(\bar{y}_T - \bar{y}_C) = Var(\bar{y}_T) + Var(\bar{y}_C)$. Consequently, $Var(\bar{y}_T - \bar{y}_C)$ in the cluster-randomized design may be estimated by $\widehat{Var}(\bar{y}_T) + \widehat{Var}(\bar{y}_C)$. In a balanced design, $\widehat{Var}(\bar{y}_T - \bar{y}_C) = Tvar / H_T + Cvar / H_C$.

### 4.1.2 Constructed Data Example

In this example, a population of 70 schools is available for the study. Thirty-five of the schools are randomly assigned to the treatment group, and the other 35 are assigned to the control group. Each school has 40 participating students. All participating students in the control schools receive the control protocol. The targeted students in each treatment school are randomly assigned to one of the four ICs within their school (e.g., a remedial pull-out program), so each IC has 10 students.

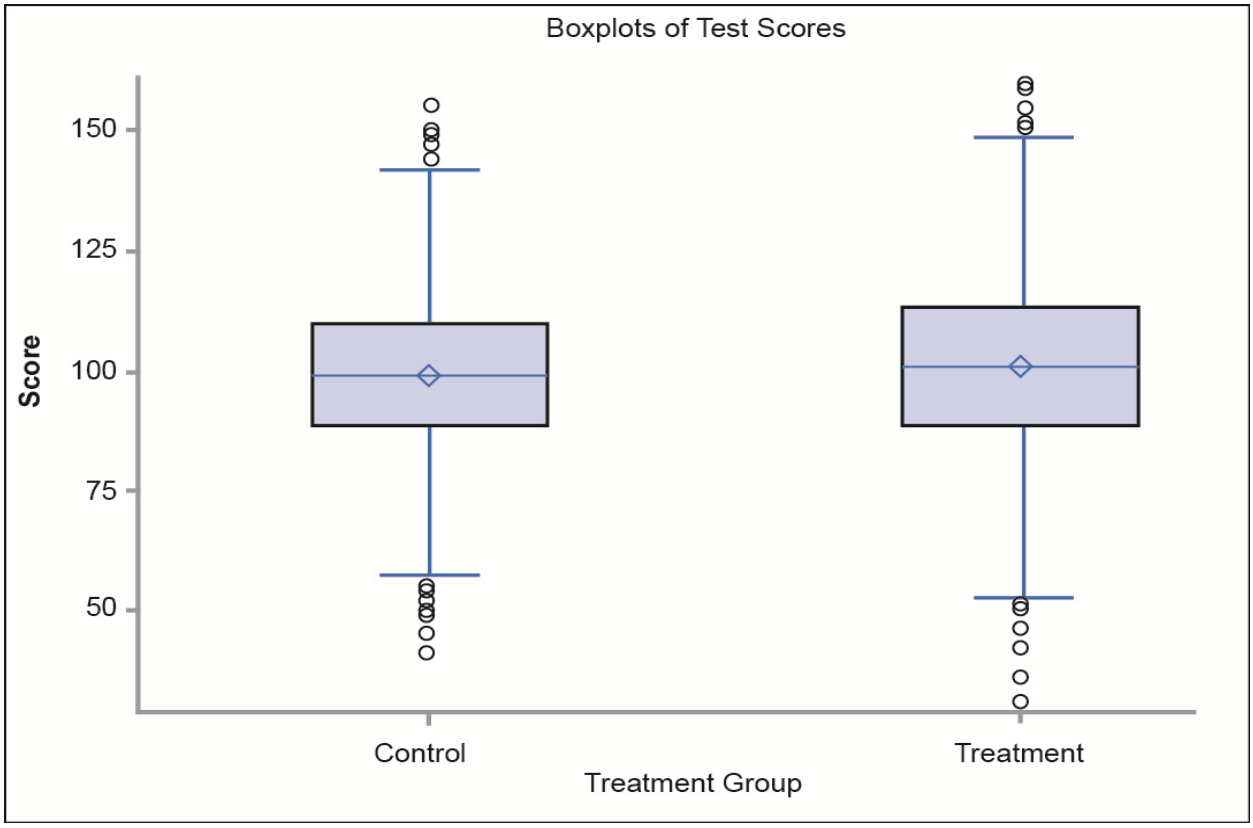
The data were generated using equation (4.1) under the assumption that the intraclass correlation coefficient is 0.15 for schools and 0.1 for ICs. The variances used for this example are: $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T} = 225 = 15^2$ for the student-level (residual) variance in each group, $\sigma^2_{\xi C} = \sigma^2_{\xi T} = 45$, and $\sigma^2_{\theta} = 30$. The data were generated with a mean of 100 scale points for students in the control group and 103 scale points for students in the treatment group. The number of schools was chosen to achieve a desired standardized minimum detectable treatment effect of 0.3 standard deviations with 80 percent power.[21]

The plots from figures 8 and 9 from section 3.3 are also appropriate for these data. Figure 15 gives the boxplot for all of the students in the control and treatment schools. The IC effects cause the treatment students to have larger within-school variability than the control students. Figure 16 displays the relative magnitudes of the within-school and between-school mean variability for the 70 schools in the study. Note the large variability among school means for both the treatment and

[21] See section 4.3 for the power calculations for this example. This corresponds to a treatment effect of 5.2 scale points. The values of the variances are determined so the intraclass correlation at the school level is 0.15, and the intraclass correlation at the IC level is 0.1. This results in $\sigma^2_{\xi C} = \sigma^2_{\varepsilon C} / 5$ and $\sigma^2_{\theta} = \sigma^2_{\varepsilon C} * 1.2 / 9$.

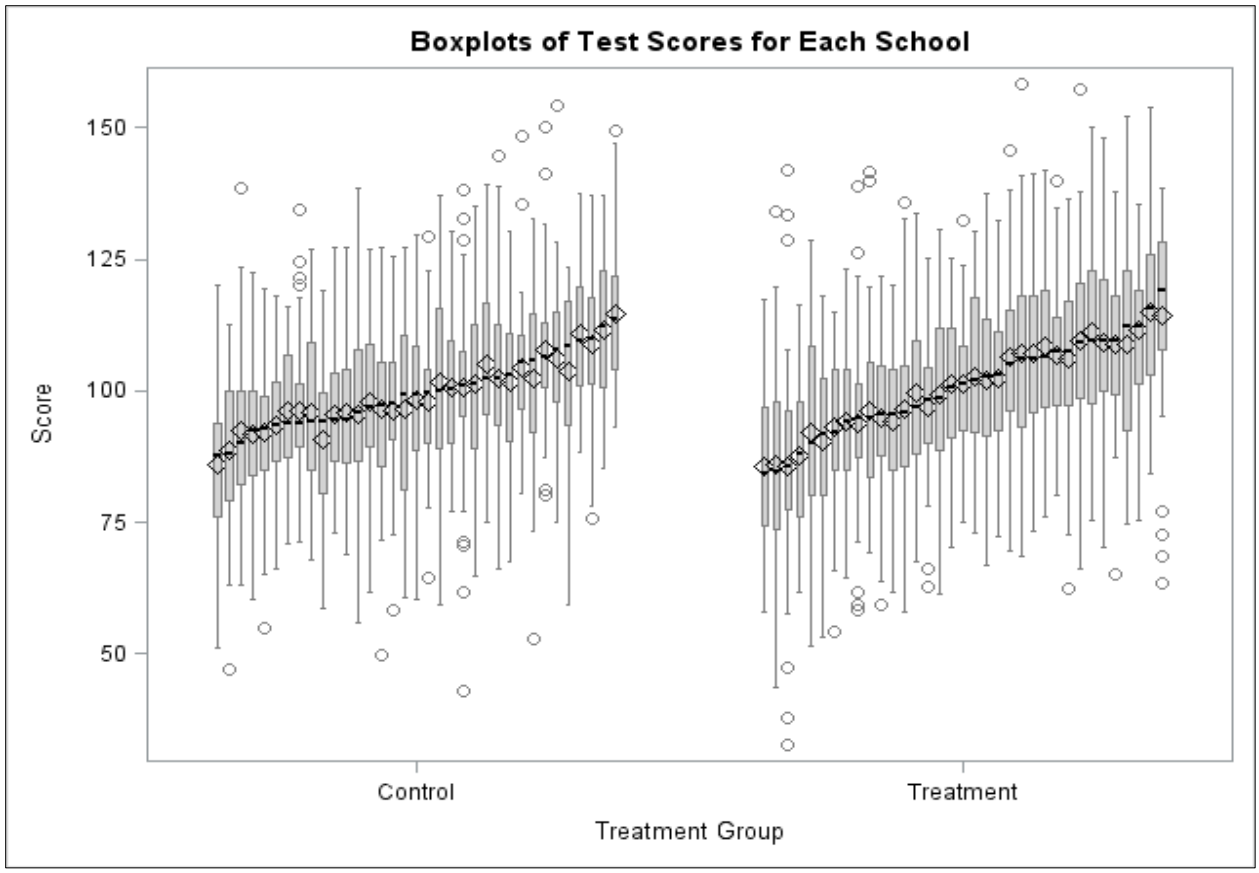
control groups. That variability reflects both the school-to-school variability and the student-to-student variability within schools (and, for the treatment schools, the IC-to-IC variability as well). For this balanced clustered design, the overall mean for the control group is the average of the 35 individual control school means, and its standard error is the sample standard deviation of the control school means divided by $\sqrt{35}$, with a similar calculation for the treatment group. No unusual data features are apparent in figure 16, but with other data sets, the plot might be used to detect unusual scores for schools and/or students. The SAS code used to obtain these plots is similar to that in figure 7.

**Figure 15. Boxplot of scores from control and treatment groups for data set in the clustered design**




**NOTE:** Symbols used in the plot are described in figure 8.

Figure 16. Plot of data, showing individual boxplots for the schools in the control and treatment groups. Within each group, the schools are arranged in order of increasing median test score




**NOTE:** Symbols used in the plot are described in figure 8.

The code in figure 17 allows the general structure in equation (4.1) to be fit in SAS. The PARMS statement and GROUP option are used to allow the variance components at all levels (school, IC, and student) to differ between the treatment and control groups. To fit a model in which these are constrained to be equal, simply omit the GROUP option. The HOLD option is again used to force the IC-level variance in the control group to equal 0. The initial values for the parameter estimates are calculated from mixed models fitted separately for the treatment and control group students.

Figure 17. SAS code to analyze data from the cluster-randomized design

```
proc mixed data=model2 noclprint;
   class trtname school ic subjid;
   model y = trt / ddfm = sat solution;
   random intercept / group=trtname subject=school(trtname) ;
   random intercept / group=trtname subject=ic(school trtname) ;
   repeated subjid / group=trtname ;
   parms (38) (58) (0) (31) (221) (233)/ hold = 3;
   title 'Random Effects Analysis for Cluster-Randomized Design';
```

We omit the information on class levels and model fit. The SAS log gives the message that the **G** matrix is not positive definite, as we expected, but otherwise indicates no convergence problems. As requested, SAS estimates six covariance parameters and two fixed effects (the intercept and slope, which equals the number of columns in X). Schools serve as the 70 subjects in this analysis because schools are the units at the top of the hierarchy.

| Dimensions | |
|---|---|
| Covariance Parameters | 6 |
| Columns in X | 2 |
| Columns in Z Per Subject | 10 |
| Subjects | 70 |
| Max Obs Per Subject | 40 |

The covariance parameter estimates are essentially the same as the input initial values. This is because SAS uses only the data in the control group to compute the control variance parameters and uses only the data in the treatment group to compute the treatment variance parameters. The IC effect for the control schools is 0, as requested. In the control group, it is estimated that about 15 percent [=38/(38+221)] of the variability comes from the variability among schools and the remaining 85 percent comes from the variability among students within the school. In the treatment group, it is estimated that about 18 percent [=59/(59+3233)] of the variability is from the school level, 10 percent [=31/(59+31+233)] from the IC-to-IC variability, and the remainder from the student-to-student variability.[22]

[22] We omit the residual analysis and regression diagnostics for space reasons; however, these analyses should always be conducted.

| Covariance Parameter Estimates | | | |
|---|---|---|---|
| Cov Parm | Subject | Group | Estimate |
| Intercept | school(trtname) | trtname Control | 38.3656 |
| Intercept | school(trtname) | trtname Treatment | 58.6294 |
| Intercept | ic(trtname*school) | trtname Control | 0 |
| Intercept | ic(trtname*school) | trtname Treatment | 30.8832 |
| subjid | | trtname Control | 221.24 |
| subjid | | trtname Treatment | 232.84 |

The estimated treatment effect is 1.42 scale points, which is not significantly different from 0 (*p*-value = 0.44). Note that the Satterthwaite method gives 64.2 degrees of freedom. This is slightly different from the 68 degrees of freedom one would expect from an analysis with 70 schools. With this large of a sample of schools, however, the degrees of freedom make no difference to the inference because a *t* distribution with 64 degrees of freedom is very similar to a *t* distribution with 68 degrees of freedom.

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard error | DF | t Value | Pr > \|t\| |
| Intercept | 99.3217 | 1.1199 | 34 | 88.69 | <.0001 |
| Trt | 1.4222 | 1.8211 | 64.2 | 0.78 | 0.4377 |

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| Trt | 1 | 64.2 | 0.61 | 0.4377 |

For the clustered design, the variability among control school means is the appropriate error term for the overall control group mean, and the variability among treatment school means is the appropriate error term for the overall treatment group mean. We may, thus, check the results of the above analysis by noting that $Var\left(\bar{y}_T - \bar{y}_C\right) = Var(\bar{y}_T) + Var(\bar{y}_C)$ and estimating the two variances separately.[23] For the control group, the sample variance among school means is 43.90; for the treatment group, the sample variance among school means is 72.18 (the larger variance in the treatment group occurs because of the IC effects). Then,

$$\widehat{Var}\left(\bar{y}_T - \bar{y}_C\right) = \frac{43.90}{35} + \frac{72.18}{35} = 3.32$$

and the test statistic is $T = 1.42 / \sqrt{3.32} = 0.78$, as obtained in the full analysis by SAS above. This result is exact for the balanced design considered here and will be approximately correct for unbalanced designs.

[23] It is important to find separate, not pooled, estimates of the variance when doing this analysis, particularly if the numbers of schools differ in the treatment and control groups. A pooled analysis, particularly with unbalanced data, can lead to incorrect *p*-values.

Finally, it is important to note that under the clustered design with school-level randomization, correlations in the data might exist due to the grouping of students in *regular* classrooms. These correlations would pertain to *both* treatment and control group students. Analytic issues surrounding the statistical treatment of these regular classroom effects are very similar to the issues discussed above for treating IC effects. In particular, the treatment of school effects as random implies that the lower level classroom effects are also random by extension. Regular classroom effects could be included as additional random errors in the estimation model in equation (4.1); the resulting variance expressions would include variance terms that represent the variability of mean outcomes across classrooms (within schools) for both research groups. Analytic complications arise, however, if IC formation cuts across regular classrooms, thus leading to cross-nested designs (see section 4.2).

### 4.1.3 Random Assignment of Schools Within Districts

The clustered design may also be replicated in different districts (or other blocking units such as cities or states). In this case, the basic design of randomly assigning entire schools to the treatment or control condition is replicated across different districts. The same structure applies to designs where classrooms within each school are randomized, and students in the treatment classrooms are then assigned to ICs.

We first examine the model for a blocked design with cluster-level randomization where district effects are treated as random. Let $d = 1$ to $D$ denote the districts available for the study. Because entire schools are randomly assigned to the treatment or control condition, the treatment indicator depends only on district and school. Let

$y_{dhij}$ =test score of student *j* in IC *i* of school *h* in district *d*
$\delta_d$ =effect of district *d*
$\xi_{dh}$ =effect of school *h* in district *d*
$\theta_{dhi}$ =effect of IC *i* in school *h* of district *d*
$\varepsilon_{dhij}$ =student-level error (residual) term for student *j* in IC *i* of school *h* in district *d*
$T_{dh}$ =treatment indicator for school *h* in district *d*: 1 for treatment and *O* for control.

We assume that $\delta_d \sim N(0,\sigma_\delta^2)$, $\xi_{dh} \sim N(0,\sigma_\xi^2)$, $\theta_{dhi} \sim N(0,\sigma_\theta^2)$, and $\varepsilon_{dhij} \sim N(0,\sigma_\varepsilon^2)$ for districts, schools, ICs, and students, respectively, and that all of those random effects are independent. If desired, different variances, $\sigma_{\varepsilon T}^2$ and $\sigma_{\varepsilon C}^2$, may be used for treatment and control group students.

In many implementations of this design, there will be relatively few districts, so estimates of a random district-by-treatment interaction effect will be unstable. Therefore, equation (4.3) omits a district-by-treatment interaction in the error term:

$$y_{dhij} = \beta_0 + \beta_1 T_{dh} + [\delta_d + \xi_{dh} + \theta_{dhi} T_{dh} + \varepsilon_{dhij}]. \quad (4.3)$$

As in the blocked PN-RCT considered in section 3.5, the district-level effects cancel in a balanced design, as shown in appendix A.

There may be instances, however, where researchers want to include a random district-by-treatment interaction term. This model would be applicable if a large number of districts participate in the study. Thus, for equation (4.4), we assume that the random district effects are independently generated from a bivariate normal distribution with variances $\sigma_\delta^2$ and $\sigma_\varphi^2$ and covariance $\sigma_{\delta\varphi}$:

$$y_{dhij} = \beta_0 + \beta_1 T_{dh} + [\delta_d + \varphi_d T_{dh} + \xi_{dh} + \theta_{dhi} T_{dh} + \varepsilon_{dhij}]. \quad (4.4)$$

The variance of the ATE for this model is given in equation (A.8) of appendix A.

We have already illustrated the main features of SAS output for this design in the previous sections, so we do not include a constructed data example for this blocked design. Sample SAS code for constructing and analyzing a data example is provided for this design in appendix E, however.

Finally, researchers may prefer to treat the district effects and their interactions with the treatment as fixed effects rather than as random effects. This might be the case if there are only a few districts in the study, or if districts are chosen to exemplify certain characteristics, such as high poverty, high numbers of Hispanic students, or other features. In that case, the school-to-school variability within the district-by-treatment interaction will form the error term for evaluating the treatment impact.

## 4.2 Cross-Nested Designs

Thus far, we have considered hierarchical (nested) experimental designs in which each lower level unit (e.g., a student) belongs to *exactly one* unit at each higher level (e.g., an IC, classroom, or school). In the design in section 4.1.3, where schools are randomized within districts, we have assumed that a student in the treatment group belongs to exactly one IC that belongs to exactly one school that in turn belongs to exactly one district. Some designs in education research, however, have a "non-nested" or "cross-nested" structure. For example, students can change primary schools during the study period, in which case students can be associated with more than one school. As another example, students in the same middle school may attend different high schools, and high schools may serve students from different middle schools. Thus, if the goal of the study is to estimate middle and high school contributions to student achievement, a cross-nested or multiple membership design structure results (see, for example, Goldstein 1995). A similar cross-nested structure occurs for studies that aim to estimate the relative contributions of neighborhoods and schools to student outcomes (Raudenbush 1993).

Adapting the literature on cross-nested designs to the estimation of treatment effects in PN-RCTs is beyond the scope of this paper. In this section, we discuss and give examples of various types of cross-nested designs in which the cross-nested structure is due solely to the IC formation. Under cross-nested designs, the model error structure becomes more complex than under fully nested designs. Although this error structure can be specified in most instances, estimation procedures using a mixed model framework become more complex and rely on additional model assumptions. The additional computational complexity occurs because the variance-covariance matrices required for estimation are no longer block diagonal (see appendix A). An example of a model that may be used to capture the covariance structure for cross-nested PN-RCT designs is given in section A.5. Computational methods for these models are a subject of ongoing research (see, for example, Cameron, Gelbach, and Miller 2011; Fielding and Goldstein 2006; and Karl, Yang, and Lohr 2013).

### 4.2.1 Changes in IC Membership Over Time

In some PN-RCTs, students may switch ICs during the follow-up period, which can lead to non-nested error structures. This may occur, for example, if a student's schedule changes or the initial IC placement is ultimately not the best fit for the student. For instance, a student may switch pull-out programs mid-year or attend after-school programs on different days. In these instances, students may be associated with multiple ICs, so students are not uniquely categorized into ICs.

Several analytic approaches can be used to reduce these designs to those considered in this paper. First, students may be coded to the IC to which they attended for the longest period or to the initial IC. If this approach is adopted, then the models described in chapter 3 and section 4.1 may be used to analyze the data. An alternative approach is to modify the basic PN-RCT model in equation (3.3) so that the treatment indicator variable may be replaced by variables describing the fraction of the time the student spent in each IC. The mathematical form for the modified model requires a different notational setup and is described in the last paragraph of section A.1 in appendix A.

### 4.2.2 Teachers are in Charge of Multiple ICs

In some PN-RCTs, the same teacher may teach more than one IC. For example, a tutor may coach more than one group of students. This can lead to correlated outcomes for treatment group members who are in different ICs but who have the same teacher. In the basic PN-RCT design, this leads to a nested design with students in the treatment group nested in ICs that are in turn nested in teachers. But consider a blocked PN-RCT design where students are randomly assigned to the treatment and control groups within schools and where the treatment students attend after-school programs (ICs) in a few central locations. In this case, the ICs are formed from students belonging to different blocks. Then, this blocked PN-RCT is an example of a "cross-nested" design because students in *different* blocks are nested within the same teacher. The analysis of such a design needs to account for the similarity among students in the same blocks (schools) as well as the extra similarity among students who attend the same IC. The resulting complex covariance structure can be specified using the methods described in Appendix A.

Cross-nested designs could also arise under the clustered design from section 4.1 (with school-level randomization) if teachers lead multiple ICs across different schools. In the evaluation of the Impacts of Comprehensive Teacher Induction (Glazerman et al. 2010), 418 elementary schools from 17 school districts serving low-income students were randomly assigned to a treatment or control group. In the treatment schools, novice teachers were offered mentoring services from experienced,

trained full-time mentors. Each mentor was assigned to an average of 12 treatment group teachers. Teachers who received intervention services from the same mentor can be considered to be in the same IC and may have had correlated outcomes because of common characteristics and intervention experiences. Mentors typically served teachers from multiple treatment schools. For example, Mentor A served teachers in Schools 1 and 2; Mentor B served teachers in Schools 1, 2, and 4; and Mentor C served teachers in schools 3, 4, and 5. Consequently, ICs are no longer nested with schools. A crossed random effects model may be used to estimate parameters in this model (see Raudenbush 1993).

### 4.2.3 ICs That Cut Across Schools or Classrooms

In some PN-RCTs, ICs may consist of students from different schools or classrooms. For example, consider the design in section 4.1 where (1) classrooms are randomized within schools and (2) the intervention is a pull-out program where ICs are formed using students across different classrooms. In this design, ICs are not fully nested within classrooms. Section A.5 of appendix A describes a model that can be used to capture the correct covariance structure for this cross-nested design.

As another example, consider an after-school intervention that serves treatment students from multiple schools. Under the designs in section 4.1 (with school-level random assignment) or 3.5 (with student-level random assignment within schools and random school effects), this grouping could lead to non-nested designs where ICs are no longer uniquely nested within schools. Similar to the approaches from above, the IC effects can be excluded from the estimation models if the higher level school effects are included in the models. It is important, however, that the treatment and control group variances are allowed to differ.

## 4.3 Statistical Power for PN-RCTs

An important part of any evaluation design is the statistical power analysis, which demonstrates how well the design of the study will be able to distinguish real impacts from chance differences. There is a growing literature on appropriate methods for conducting power analyses in education RCTs (see, for example, Hedges and Rhoads 2010; Jenney and Lohr 2009; and Schochet 2008), but no comparable literature for PN-RCT and related designs where ICs are formed in the treatment group. This section aims to close this gap by building on the previous literature to consider statistical power issues for the various designs considered in chapter 3 and section 4.1. For lack of a better label, we

refer to all designs as PN-RCTs even though, technically, the designs in section 4.1 are C-RCTs with clustering in both study arms.

The presence of IC effects in PN-RCTs complicates the power analysis because the variance expressions differ for the treatment and control groups, with the treatment group having a source of variability not found in the control group. Furthermore, because the variance structure differs in the control and treatment arms, it may be desirable to have more students assigned to the research condition with the larger variance (which will often be the treatment group). The power calculations given in this section consider the possibility of unequal variations and allocations.

Note that for all designs considered, a simple but conservative power analysis can be performed by assuming that the larger treatment-group variance also applies to the control group. For the basic PN-RCT design, this conservative analysis would be done by pretending that the control group also has variability due to ICs and calculating the power using C-RCT formulas. Since the actual PN-RCT variance is smaller than assumed in this analysis, this approach will give a conservatively large sample size for the experiment.

This section is in three parts. First, we define statistical power and then provide an overview of our theoretical framework for the sample size calculations. Finally, we provide sample size formulas and present illustrative sample size calculations using empirically based parameter values. Our analysis pertains to both small- or larger scale PN-RCT designs and to interventions with small or large effects.

### 4.3.1 Defining Minimum Detectable Impacts

To determine appropriate sample sizes for impact evaluations, researchers typically calculate minimum detectable impacts, which represent the smallest program impacts—average treatment and comparison group differences—that can be detected with a high probability. In addition, it is common to standardize minimum detectable impacts into *effect size units*—that is, as a percentage of the standard deviation of the outcome measures (also known as Cohen's *d*)—to facilitate the comparison of findings across outcomes that are measured on different scales (Cohen 1988). Hereafter, minimum detectable impacts in effect size units are denoted as "*MDEs*."

Mathematically, the $MDE$ formula can be expressed as follows:

$$MDE = Factor(\alpha, \beta, df)\sqrt{Var(Impact)} / \sigma \,, \tag{4.5}$$

where $Var(Impact)$ is the variance of the impact estimate, $\sigma$ is the standard deviation of the outcome measure, and $Factor(\alpha, \beta, df)$ is a constant that is a function of the significance level ($\alpha$), statistical power ($\beta$), and the number of degrees of freedom ($df$) that was discussed in chapter 3.[24] $Factor(.)$ becomes larger as $\alpha$ and $df$ decrease and as $\beta$ increases (see table 4).

A key issue for any RCT is the precision standard to adopt for the impact estimates. This decision determines appropriate sample sizes because higher precision standards will require larger samples. Two key factors need to be considered. First, the precision standard should depend on what impact is deemed to have practical significance in terms of future, longer term student outcomes. Second, the precision standard should depend on what intervention effects are realistically attainable. Recent research has discussed several approaches for selecting appropriate $MDE$ targets for education evaluations using such frameworks as typical (normative) growth in student outcomes during a school year, differences between subgroups of students and schools with recognized practical significance, the effects found for other similar interventions, and cost (Bloom, Richburg-Hayes, and Black 2008; Lipsey et al. 2012; Schochet 2008). Precision targets for a particular evaluation will depend critically on the characteristics of the students included in the study (such as their ages and achievement levels), the intensity of the interventions, and the nature of the primary outcome measures (e.g., proximal or mediator outcomes may require smaller sample sizes than more distal student achievement outcomes).

[24] Specifically, $Factor(.)$ can be expressed as $[T^{-1}(\alpha) + T^{-1}(\beta)]$ for a one-tailed test and $[T^{-1}(\alpha/2) + T^{-1}(\beta)]$ for a two-tailed test, where $T^{-1}(.)$ is the inverse of the student's *t* distribution function with *df* degrees of freedom.

Table 4. Values for $Factor(.)$ in equation (4.5) of the text, by the number of degrees of freedom, for one- and two-tailed tests, and at 80 and 85 percent power

| Number of degrees of freedom | One-tailed test | | Two-tailed test | |
|---|---|---|---|---|
| | 80 Percent power | 85 Percent power | 80 Percent power | 85 Percent power |
| 2 | 3.98 | 4.31 | 5.36 | 5.69 |
| 3 | 3.33 | 3.61 | 4.16 | 4.43 |
| 4 | 3.07 | 3.32 | 3.72 | 3.97 |
| 5 | 2.94 | 3.17 | 3.49 | 3.73 |
| 6 | 2.85 | 3.08 | 3.35 | 3.58 |
| 7 | 2.79 | 3.02 | 3.26 | 3.49 |
| 8 | 2.75 | 2.97 | 3.20 | 3.42 |
| 9 | 2.72 | 2.93 | 3.15 | 3.36 |
| 10 | 2.69 | 2.91 | 3.11 | 3.32 |
| 11 | 2.67 | 2.88 | 3.08 | 3.29 |
| 12 | 2.66 | 2.87 | 3.05 | 3.26 |
| 13 | 2.64 | 2.85 | 3.03 | 3.24 |
| 14 | 2.63 | 2.84 | 3.01 | 3.22 |
| 15 | 2.62 | 2.83 | 3.00 | 3.21 |
| 20 | 2.59 | 2.79 | 2.95 | 3.15 |
| 30 | 2.55 | 2.75 | 2.90 | 3.10 |
| 40 | 2.54 | 2.74 | 2.87 | 3.07 |
| 50 | 2.53 | 2.72 | 2.86 | 3.06 |
| 100 | 2.51 | 2.70 | 2.83 | 3.03 |
| Infinity | 2.49 | 2.68 | 2.80 | 3.00 |

**NOTE:** All figures assume a 5 percent significance level.

### 4.3.2 Overview of Theoretical Approach

Treatment and control group student sample sizes, $n_T$ and $n_C$, enter equation (4.5) as part of $Var(Impact)$. Thus, for a given design structure, equation (4.5) can be used to solve for $n_T$ and $n_C$ to attain a target $MDE$. This calculation is more complex for PN-RCTs than for typical RCTs because the variance structure for PN-RCTs differs for the treatment and control groups, whereas the power analysis literature for RCTs typically assumes equal variances for the two research groups.

We use the following staged approach to obtain mathematical formulas for sample size calculations for the PN-RCT designs considered in chapter 3 and section 4.1:

- **Develop sample size formulas for a "reference" design with no hierarchical IC structure.** Our reference design is the basic PN-RCT with student-level randomization where IC effects are treated as fixed; this is an I-RCT design.
- **Develop "design effect" formulas for other PN-RCTs relative to the reference design.** The design effect is the *ratio of the variances of the impact estimates* of the considered

PN-RCT design relative to the reference design. These design effects will typically be greater than 1.

- **For each PN-RCT design, calculate sample sizes to attain a given *MDE* value by multiplying the required sample sizes under the reference design (from Step 1) by the design effect (from Step 2).** For example, if 150 treatment and 150 control students are required to obtain an *MDE* of .3 standard deviations under the reference design and the design effect for a particular PN-RCT design is 1.4, the corresponding sample size for that PN-RCT would be 210 treatments and 210 controls.

In what follows, we discuss these steps in more detail using the variance formulas and notation from chapter 3 and section 4.1.

### 4.3.3 The Reference Design: An I-RCT Design

The reference design for our analysis is the basic PN-RCT design with student-level random assignment, where IC effects are treated as fixed. We adopt this design because it is the most common design specification used in education research for non-clustered RCT designs and applies to both blocked and unblocked designs. The variance of the treatment effect under this design (without the inclusion of baseline model covariates) has the simple form:

$$Var\left(Impact\right) = \frac{\sigma_\theta^2 + \sigma_{\varepsilon T}^2}{np} + \frac{\sigma_{\varepsilon C}^2}{n(1-p)}, \tag{4.6}$$

where $p$ is the proportion of students who are randomly assigned to the treatment group, $n = n_T + n_C$ is the total student sample size, and other parameters are defined as in chapter 3 and section 4.1.

During the design stage of an evaluation, researchers may not have specific information on the extent to which student-level variances will differ for treatment and control students. Thus, in what follows, we assume that $\sigma_{\varepsilon T}^2 = \sigma_{\varepsilon C}^2 = \sigma_\varepsilon^2$. This specification assumes that the treatment group variance will be *larger* than the control group variance (which might not always hold in practice). It assumes also that the treatment group variance will be influenced by the heterogeneity of IC effects but not by the heterogeneity of treatment effects from other sources.

To generalize equation (4.6) to allow for the inclusion of baseline covariates (e.g., pretests) and block indicators, we assume that the covariates will reduce each variance component by a common factor

$\left(1-R^2\right)$. Thus, under our simplifying assumptions, we can express the variance for the reference design with baseline covariates as:

$$Var\left(Impact\right)=\left[\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{np}+\frac{\sigma_\varepsilon^2}{n\left(1-p\right)}\right]\left(1-R^2\right). \tag{4.7}$$

We can then use equations (4.5) and (4.7) to solve for $n$ to ensure that the reference design has a high probability (say, 80 percent) of detecting a true standardized impact equal to a pre-specified $MDE$ value. An important analytic issue is what value of $\sigma$ to use in equation (4.5) to standardize the impacts into effect size units. A common approach, which we adopt, is to use the control group (status quo) standard deviation, which is $\sigma_\varepsilon$ for our reference design. An alternative approach is to use the full sample standard deviation $\sqrt{p\sigma_\theta^2+\sigma_\varepsilon^2}$, which uses the weighted average of the treatment and control group variances.

Using this approach, we find that the sample size formula for the reference design is:

$$n_{Ref}=\left[\frac{1}{p\left(1-p\right)}\right]\left[\frac{\left(1-R^2\right)Factor(.)^2}{MDE^2}\right]\left[\frac{\left(1-p\rho_\theta\right)}{\left(1-\rho_\theta\right)}\right], \tag{4.8}$$

where $\rho_\theta=\sigma_\theta^2/(\sigma_\theta^2+\sigma_\varepsilon^2)$ is the intraclass correlation coefficient (ICC) at the IC level for the treatment group. Note that if $\rho_\theta$ is set to 0, we obtain the usual sample size formula found in the literature for non-clustered RCT designs. The associated treatment and control student sample sizes can be calculated using $n_{Ref,T}=n_{Ref}p$ and $n_{Ref,C}=n_{Ref}(1-p)$, respectively.

Table 5 displays illustrative total sample size calculations for $n_{Ref}$ using equation (4.8) assuming a 5 percent significance level for a two-tailed test at 80 percent power and equal treatment and control group sample sizes ($p=.5$). These are standard assumptions used in power calculations in the education field. To provide a range of realistic sample sizes that can be used in practice, we perform the calculations assuming various empirically based values for $MDE$, $R^2$, and $\rho_\theta$. Using results found in Schochet (2008); Bloom et al. (2007); Hedges and Hedberg (2007); and Raudenbush, Martinez, and Spybrook (2007), we allow (1) $MDE$ values to range from .10 to .50, which are impact values typically found for promising interventions in the education field and often used as targets in power calculations for education RCTs, and (2) $R^2$ values to range from 0 (for models without baseline covariates) to .75 (for models that include highly predictive baseline pretest scores).

There is less literature on plausible values for $\rho_\theta$ (which is an important area for future research). Thus, we use values for $\rho_\theta$ that range from 0 to .2 based on intraclass correlations at the *classroom* level that have been found in the literature (see, for example, Schochet 2008).

**Table 5. Total sample size calculations for students for the reference design (the basic PN-RCT with student randomization and fixed IC effects), for treatment and control groups of equal size**

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Intraclass correlation ($\rho_\theta$) | 0 | .25 | .50 | .75 |
| | MDE Target = .10 | | | |
| 0.0 | 3,140 | 2,355 | 1,570 | 785 |
| 0.1 | 3,315 | 2,486 | 1,657 | 829 |
| 0.2 | 3,533 | 2,650 | 1,767 | 883 |
| | MDE Target = .20 | | | |
| 0.0 | 785 | 589 | 393 | 196 |
| 0.1 | 829 | 622 | 414 | 207 |
| 0.2 | 883 | 662 | 442 | 221 |
| | MDE Target = .30 | | | |
| 0.0 | 349 | 262 | 174 | 87 |
| 0.1 | 368 | 276 | 184 | 92 |
| 0.2 | 393 | 294 | 196 | 98 |
| | MDE Target = .40 | | | |
| 0.0 | 196 | 147 | 98 | 49 |
| 0.1 | 207 | 155 | 104 | 52 |
| 0.2 | 221 | 166 | 110 | 55 |
| | MDE Target = .50 | | | |
| 0.0 | 126 | 94 | 63 | 31 |
| 0.1 | 133 | 99 | 66 | 33 |
| 0.2 | 141 | 106 | 71 | 35 |

**NOTES:** All calculations were conducted using equation (4.8) in the text assuming a 5 percent significance level, two-tailed test at 80 percent power and equal treatment and control group sample sizes (that is, p=.5). For simplicity, all calculations assume $Factor(.) = 2.802$ regardless of the degrees of freedom. The figures in the table show total student sample sizes (split evenly between the treatment and control groups) that are required to achieve the indicted precision targets measured in effect size units. The figures are shown for various MDE targets, IC-level ICCs ($\rho_\theta$), and regression $R^2$ values.

As an example of how to interpret the entries in table 5, we find that the required total sample size, $n_{Ref}$, is 622 students (311 treatments and 311 controls) if $MDE = .20$, $\rho_\theta = .10$ and $R^2 = .25$. Under these assumptions, the required sample size reduces somewhat from 622 to 589 students under a traditional RCT design without ICs (where $\rho_\theta = 0$). Consistent with results from the literature, required sample sizes become smaller as $R^2$ and $MDE$ values increase and as $\rho_\theta$ values

decrease. $R^2$ values have a particularly large effect on precision; thus, the collection of detailed baseline variables (and especially, pretests) is an important design feature for improving the precision of the impact estimates.

Note that in traditional RCTs that assume equal treatment and control group variances, for a given sample size, an equal treatment-control split with $p = .5$ yields the most precise impact estimates. However, in PN-RCTs where the treatment and control group variances could differ, the optimal treatment-control allocation will place more students into the research group with the larger variance. For the reference design in equation (4.7), the variance of the impact estimates will be minimized if

$$p_{Opt} = \frac{1}{1+\sqrt{1-\rho_\theta}}. \tag{4.9}$$

Thus, for example, if $\rho_\theta = .2.$, we find that $p_{Opt} = .53$, or that the most precise impact estimates would be obtained if 53 percent of students were randomized to the treatment group and 47 percent were randomized to the control group. For most realistic values of $\rho_\theta$, however, the optimal treatment-control split does not stray too far from 50-50, and thus, we do not consider this issue further below for the other designs. Furthermore, there are also other advantages to maintaining a 50-50 design, such as ease of administration and simplified recruitment. In addition, in many experiments, the treatment protocol will be more expensive to implement than the control protocol will, so the optimal allocation in equation (4.9) would be moderated by cost considerations. Taking costs into account, we derive the following expression for the optimal allocation:

$$p_{Opt} = \frac{1}{1+\sqrt{(costs_T / costs_C)(1-\rho_\theta)}}, \tag{4.10}$$

where $costs_T$ and $costs_C$ are study costs per treatment and control group student, respectively.

### 4.3.4 The Basic PN-RCT: Student Randomization; Random IC Effects

In this section, we consider the basic PN-RCT design discussed in chapter 3, where IC effects for the treatment group are treated as random effects. These random IC effects will reduce the precision of the impact estimates. Using the notation of chapter 3, we can write the variance of an impact estimate for this random effects design as

$$Var(Impact) = \left[ \frac{\sigma_\varepsilon^2}{n_T} + \frac{\sigma_\theta^2}{I_T} + \frac{\sigma_\varepsilon^2}{n_C} \right] \left(1 - R^2\right), \tag{4.11}$$

where $I_T$ is the number of ICs and where other parameters are defined as above. The critical difference between this design and the reference design is that $\sigma_\theta^2$ is divided by $I_T$ rather than by $n_T$. It is assumed in equation (4.11) that the inclusion of baseline covariates in the model reduces each variance component by the same factor $\left(1 - R^2\right)$, an assumption that we make hereafter.

The design effect for the random effects design can be calculated by dividing the variance in equation (4.11) by the variance of the reference design in equation (4.7) to yield

$$deff = 1 + \frac{\rho_\theta \left(1 - p\right)(J - 1)}{(1 - \rho_\theta p)}, \tag{4.12}$$

where $J = n_T / I_T$ is the (average) number of treatment students per IC and where other parameters are defined as above. A critical feature of this expression is that the design effect increases as the IC sample size increases. Stated differently, for a given total sample size, we obtain more precise impact estimates if we increase the number of sampled ICs and decrease the number of students sampled per IC. If $J$ equals 1, this PN-RCT design reduces to the reference design from above because the design effect is 1 in this case.

To achieve a target $MDE$ value, we can now calculate required student samples for the random effects design by multiplying the design effect in equation (4.12) by the corresponding sample size for the reference design in table 5. Table 6 displays such illustrative sample sizes, where we allow $J$ to range from 2 to 20 to allow for designs with small ICs (e.g., pull-out groups) and larger ICs (e.g., entire classrooms). The structure of the table is similar to table 5 except that we omit some rows to keep the presentation manageable.

As an example to show how the figures in table 6 were calculated, note first using equation (4.12) that if $\rho_\theta = .2$, $p = .5$, and $J = 5$, the design effect is 1.44. Thus, as shown in table 6, to achieve an $MDE$ value of .2 with an $R^2$ value of .25, the required sample for the random effects design is 957 students (662 students from table 5 multiplied by the 1.44 design effect). Stated differently, the random effects design requires a 44 percent larger sample to obtain impact estimates with the same precision level as the reference design. For $J = 2$, the design effect reduces substantially from 1.44 to 1.11, and the required sample size reduces from 957 to 736 students.

**Table 6. Total sample size calculations for students for the basic PN-RCT design with random IC effects, for treatment and control groups of equal size**

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Average IC sample size | 0 | .25 | .50 | .75 |
| MDE Target = .10; $\rho_\theta$ = .1 | | | | |
| 2 | 3,489 | 2,617 | 1,745 | 872 |
| 5 | 4,013 | 3,010 | 2,006 | 1,003 |
| 10 | 4,885 | 3,664 | 2,443 | 1,221 |
| 20 | 6,630 | 4,972 | 3,315 | 1,657 |
| MDE Target = .10; $\rho_\theta$ = .2 | | | | |
| 2 | 3,926 | 2,944 | 1,963 | 981 |
| 5 | 5,103 | 3,827 | 2,552 | 1,276 |
| 10 | 7,066 | 5,300 | 3,533 | 1,767 |
| 20 | 10,992 | 8,244 | 5,496 | 2,748 |
| MDE Target = .20; $\rho_\theta$ = .1 | | | | |
| 2 | 872 | 654 | 436 | 218 |
| 5 | 1,003 | 752 | 502 | 251 |
| 10 | 1,221 | 916 | 611 | 305 |
| 20 | 1,657 | 1,243 | 829 | 414 |
| MDE Target = .20; $\rho_\theta$ = .2 | | | | |
| 2 | 981 | 736 | 491 | 245 |
| 5 | 1,276 | 957 | 638 | 319 |
| 10 | 1,767 | 1,325 | 883 | 442 |
| 20 | 2,748 | 2,061 | 1,374 | 687 |
| MDE Target = .30; $\rho_\theta$ = .1 | | | | |
| 2 | 388 | 291 | 194 | 97 |
| 5 | 446 | 334 | 223 | 111 |
| 10 | 543 | 407 | 271 | 136 |
| 20 | 737 | 552 | 368 | 184 |
| MDE Target = .30; $\rho_\theta$ = .2 | | | | |
| 2 | 436 | 327 | 218 | 109 |
| 5 | 567 | 425 | 284 | 142 |
| 10 | 785 | 589 | 393 | 196 |
| 20 | 1,221 | 916 | 611 | 305 |

**Table 6. Total sample size calculations for students for the basic PN-RCT design with random IC effects, for treatment and control groups of equal size (Continued)**

| Regression $R^2$ value from model covariates | | | | |
|---|---|---|---|---|
| Average IC sample size | 0 | .25 | .50 | .75 |
| MDE Target = .40; $\rho_\theta$ = .1 | | | | |
| 2 | 218 | 164 | 109 | 55 |
| 5 | 251 | 188 | 125 | 63 |
| 10 | 305 | 229 | 153 | 76 |
| 20 | 414 | 311 | 207 | 104 |
| MDE Target = .40; $\rho_\theta$ = .2 | | | | |
| 2 | 245 | 184 | 123 | 61 |
| 5 | 319 | 239 | 159 | 80 |
| 10 | 442 | 331 | 221 | 110 |
| 20 | 687 | 515 | 343 | 172 |
| MDE Target = .50; $\rho_\theta$ = .1 | | | | |
| 2 | 140 | 105 | 70 | 35 |
| 5 | 161 | 120 | 80 | 40 |
| 10 | 195 | 147 | 98 | 49 |
| 20 | 265 | 199 | 133 | 66 |
| MDE Target = .50; $\rho_\theta$ = .2 | | | | |
| 2 | 157 | 118 | 79 | 39 |
| 5 | 204 | 153 | 102 | 51 |
| 10 | 283 | 212 | 141 | 71 |
| 20 | 440 | 330 | 220 | 110 |

**NOTES:** All calculations were conducted assuming a 5 percent significance level, two-tailed test at 80 percent power, and equal treatment and control group sample sizes. The figures in the table show total sample sizes (split evenly between the treatment and control groups) that are required to achieve the indicted ATE precision targets measured in effect size units. The figures are shown for various precision targets, intraclass correlations $(\rho_\theta)$ and regression $R^2$ values. See the text for formulas.

The sample sizes in table 6 are all larger than the corresponding figures in table 5 because of the presence of design effects, which range from about 1.05 (when $J = 2$) to 1.30 (when $J = 5$) to 2.0 (when $J = 10$) to 2.5 (when $J = 20$). This highlights the important point that precision levels in PN-RCTs can typically be improved if more ICs and fewer students per IC are sampled for the study, to the extent that study resources allow.

### 4.3.5 C-RCT with ICs in the Treatment Group: School Randomization; Random School and IC Effects

In this section, we consider clustered designs where groups, such as schools or classrooms, rather than students are the unit of random assignment. Students in the treatment clusters are subsequently

assigned to ICs. To focus the analysis, we consider designs where groups are schools but where the analysis applies equally to classroom-based designs.

Using the notation from section 4.1, the variance expression for an impact estimate under this design is

$$Var\left(Impact\right)=\left[\frac{\sigma_{\xi}^{2}}{H_{T}}+\frac{\sigma_{\theta}^{2}}{H_{T}\times I_{TH}}+\frac{\sigma_{\varepsilon}^{2}}{n_{T}}\right]+\left[\frac{\sigma_{\xi}^{2}}{H_{C}}+\frac{\sigma_{\varepsilon}^{2}}{n_{C}}\right], \tag{4.13}$$

where $\sigma_{\xi}^{2}$ is the variance of the random school effect (which, for simplicity, is assumed to be the same for treatment and control schools). $H_T$ and $H_C$ are the number of treatment and control schools, respectively, where $H = H_T + H_C$ is the total number of schools, $I_{TH}$ is the average number of ICs per treatment school, and other parameters are defined as above. The leading variance terms for *both* the treatment and control groups are the school-level variances, which are divided by the number of schools (not the number of students); thus, random IC effects will typically have less of an influence on the variance estimates for this design than the basic PN-RCT design.

The design effect for this C-RCT design relative to the reference design can be expressed as follows:

$$deff = 1+\frac{\rho_{\xi T}\left(m_{H}-1\right)+\rho_{\theta T}\left(1-p_{H}\right)\left(J-1\right)}{\left(1-\rho_{\theta T}p_{H}\right)} \tag{4.14}$$

where

$m_H$ is the average number of students per school and equals $I_{TH}\times J$ in the treatment schools

$J$ is the average number of treatment students per IC

$p_H = H_T / H$ is the proportion of schools randomly assigned to the treatment group

$\rho_{\xi T}=\dfrac{\sigma_{\xi}^{2}}{\sigma_{\xi}^{2}+\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2}}$ is the school-level ICC for the treatment group

$\rho_{\theta T}=\dfrac{\sigma_{\theta}^{2}}{\sigma_{\xi}^{2}+\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2}}$ is the IC-level ICC for the treatment group.[25]

[25] The design effect in equation (4.14) is calculated by dividing equation (4.13) by the variance for the reference design that includes the school-level error component: $\{\sigma_{\xi}^{2}+\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2}\}/n_{T}+\{\sigma_{\xi}^{2}+\sigma_{\varepsilon}^{2}\}/n_{C}$. The control group ICC, $\rho_{\xi C}=\sigma_{\xi}^{2}/\{\sigma_{\xi}^{2}+\sigma_{\varepsilon}^{2}\}$, is embedded in this formula.

For a given total sample size of students, the design effect in equation (4.14) becomes larger as we increase the number of students per school ($m_H$) and the number of students per IC ($J$). Thus, precision levels can always be improved if more schools and fewer students per school are sampled for the study, and similarly for ICs. The design effect for the C-RCT designs reduces to the design effect for the basic PN-RCT if $\rho_{\xi T} = 0$ and $p_H = p$. Stated differently, the two designs will coincide if there are no systematic differences in average test scores across schools.

Table 7 shows the number of *schools* ($H$) that is required to attain a given $MDE$ value, where we vary the average number of students per school ($m_H$ = 5, 20, or 40) and the average number of students per IC ($J$ = 2, 5, 10, or 20). To keep the presentation manageable, we assume for all calculations that $\rho_{\xi T} = .15$ based on empirical values found in the literature (see, for example, Schochet 2008). For most calculations, we assume $\rho_{\theta T} = .10$ but also present figures where $\rho_{\theta T} = 0$ to examine the extent to which IC effects influence the sample size calculations. Sample sizes for *students* can be obtained from the figures in table 7 by multiplying the displayed sample sizes for *schools* by the number of students per school ($m_H$).

As an example of how to interpret the figures in table 7, we find that 62 schools (31 treatment and 31 control schools) would be required under the C-RCT design to attain an $MDE$ of .30 standard deviations assuming $m_H$ =5, $J = 2$, and $R^2$ =.50. The total student sample size for this design would be 310 students, compared to 194 students under the basic PN-RCT design with random IC effects, and 184 students under the reference design. In a traditional RCT without ICs ($\rho_{\theta T} = 0$), the comparable required number of schools would reduce from 62 to 56 schools, suggesting that IC effects have some influence on power. The design effects for the figures in table 5 are about 1.8 when $m_H = 5$, about 4.2 when $m_H = 20$, and about 7.4 when $m_H = 40$.

Table 7. Total sample size calculations for schools for the C-RCT design with random school and IC effects, for treatment and control groups of equal size

| Regression $R^2$ value from model covariates | | | | |
|---|---|---|---|---|
| Average IC sample size ( $J$ ) | 0 | .25 | .50 | .75 |
| **MDE Target = .10;** $m_H$ = 5* | | | | |
| 1; $\rho_{\theta T} = 0$ | 1,005 | 754 | 502 | 251 |
| 2 | 1,117 | 837 | 558 | 279 |
| **MDE Target = .10;** $m_H$ = 20* | | | | |
| 1; $\rho_{\theta T} = 0$ | 605 | 453 | 302 | 151 |
| 2 | 672 | 504 | 336 | 168 |
| 5 | 698 | 523 | 249 | 174 |
| 10 | 742 | 556 | 371 | 185 |
| **MDE Target = .10;** $m_H$ = 40* | | | | |
| 1; $\rho_{\theta T} = 0$ | 538 | 403 | 269 | 134 |
| 2 | 598 | 448 | 299 | 149 |
| 5 | 611 | 458 | 305 | 153 |
| 10 | 632 | 474 | 316 | 158 |
| 20 | 676 | 507 | 338 | 169 |
| **MDE Target = .20;** $m_H$ = 5* | | | | |
| 1; $\rho_{\theta T} = 0$ | 251 | 188 | 126 | 63 |
| 2 | 279 | 209 | 140 | 70 |
| **MDE Target = .20;** $m_H$ = 20* | | | | |
| 1; $\rho_{\theta T} = 0$ | 151 | 113 | 76 | 38 |
| 2 | 168 | 126 | 84 | 42 |
| 5 | 174 | 131 | 87 | 44 |
| 10 | 185 | 139 | 93 | 46 |
| **MDE Target = .20;** $m_H$ = 40* | | | | |
| 1; $\rho_{\theta T} = 0$ | 134 | 101 | 67 | 34 |
| 2 | 149 | 112 | 75 | 37 |
| 5 | 153 | 114 | 76 | 38 |
| 10 | 158 | 119 | 79 | 40 |
| 20 | 169 | 127 | 85 | 42 |
| **MDE Target = .30;** $m_H$ = 5* | | | | |
| 1; $\rho_{\theta T} = 0$ | 112 | 84 | 56 | 28 |
| 2 | 124 | 93 | 62 | 31 |

**Table 7. Total sample size calculations for schools for the C-RCT design with random school and IC effects, for treatment and control groups of equal size (Continued)**

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Average IC sample size ( $J$ ) | 0 | .25 | .50 | .75 |
| **MDE Target = .30; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 67 | 50 | 34 | 17 |
| 2 | 75 | 56 | 37 | 19 |
| 5 | 78 | 58 | 39 | 19 |
| 10 | 82 | 62 | 41 | 21 |
| **MDE Target = .30; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 60 | 45 | 30 | 15 |
| 2 | 66 | 50 | 33 | 17 |
| 5 | 68 | 51 | 34 | 17 |
| 10 | 70 | 53 | 35 | 18 |
| 20 | 75 | 56 | 38 | 19 |
| **MDE Target = .40; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 63 | 47 | 31 | 16 |
| 2 | 70 | 52 | 35 | 17 |
| **MDE Target = .40; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 38 | 28 | 19 | 9 |
| 2 | 42 | 31 | 21 | 10 |
| 5 | 44 | 33 | 22 | 11 |
| 10 | 46 | 35 | 23 | 12 |
| **MDE Target = .40; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 34 | 25 | 17 | 8 |
| 2 | 37 | 28 | 19 | 9 |
| 5 | 38 | 29 | 19 | 10 |
| 10 | 40 | 30 | 20 | 10 |
| 20 | 42 | 32 | 21 | 11 |
| **MDE Target = .50; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 40 | 30 | 20 | 10 |
| 2 | 45 | 33 | 22 | 11 |
| **MDE Target = .50; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 24 | 18 | 12 | 6 |
| 2 | 27 | 20 | 13 | 7 |
| 5 | 28 | 21 | 14 | 7 |
| 10 | 30 | 22 | 15 | 7 |

**Table 7. Total sample size calculations for schools for the C-RCT design with random school and IC effects, for treatment and control groups of equal size (Continued)**

| Regression $R^2$ value from model covariates | | | | |
|---|---|---|---|---|
| Average IC sample size ( $J$ ) | 0 | .25 | .50 | .75 |
| MDE Target = .50; $m_H$ = 40* | | | | |
| 1; $\rho_{\theta T}$ = 0 | 22 | 16 | 11 | 5 |
| 2 | 24 | 18 | 12 | 6 |
| 5 | 24 | 18 | 12 | 6 |
| 10 | 25 | 19 | 13 | 6 |
| 20 | 27 | 20 | 14 | 7 |

**NOTES:** All calculations were conducted using equations (4.9) and (4.14) assuming a 5 percent significance level, two-tailed test at 80 percent power, $\rho_{\xi T} = .15$, $\rho_{\theta T} = .10$ (except where otherwise noted), and equal treatment and control group school sample sizes (that is, $p_H = .5$). For simplicity, all calculations assume $Factor(.) = 2.802$ regardless of the degrees of freedom. The figures in the table show total sample sizes for *schools* (split evenly between the treatment and control groups) that are required to achieve the indicted precision targets measured in effect size units. The figures are shown for various MDE targets, number of students per school $(m_H)$, the average IC sample size $J$, school-level ICCs ( $\rho_{\xi T}$ ), IC-level ICCs ( $\rho_{\theta T}$ ), and regression $R^2$ values.

* Sample sizes for *students* can be obtained by multiplying sample sizes for *schools* by the number of students per school ( $m_H$ ).

### 4.3.6 Blocked Designs: Randomization of Students or Schools Within Blocks

Under the blocked PN-RCT design, students are randomly assigned to the treatment or control groups within schools (or school districts). For this design, the school (block) effects can be treated as either fixed or random. If the school effects are treated as fixed, the model could include school membership indicator variables as covariates. Thus, the sample size requirements for this specification are very similar to the random effects version of the basic PN-RCT if IC effects are treated as random, where the $R^2$ values are adjusted to reflect the predictive power of the school indicator variables.

If both the school and IC effects are treated as random in the blocked PN-RCT design, we can use the notation from section 3.5 to express the variance of an impact estimate as follows:

$$Var\left(Impact\right)=\left[\frac{\sigma_{\eta}^{2}}{H}+\frac{\sigma_{\theta}^{2}}{H\times p_{S}\times I_{TH}}+\frac{\sigma_{\varepsilon}^{2}}{n_{T}}+\frac{\sigma_{\varepsilon}^{2}}{n_{C}}\right], \tag{4.15}$$

where $\sigma_{\eta}^{2}$ is the variance of the *impacts* across schools, $p_S$ is the average proportion of students assigned to the treatment group within each school, and other parameters are defined as above. The design effect for this PN-RCT design is

$$deff=1+\frac{\rho_{\xi T}\left(\omega m_{H}p_{S}(1-p_{S})-1\right)+\rho_{\theta T}\left(1-p_{S}\right)(J-1)}{(1-\rho_{\theta T}p_{S})}, \tag{4.16}$$

where $\omega=\sigma_{\eta}^{2}/\sigma_{\xi}^{2}$ is the ratio of the variation in *test score impacts* across schools to the variation in *mean test scores* across schools. In most instances, $\omega$ will be less than 1, so impact estimates will typically be more precise under this design than the C-RCT design. If $\omega$ equals 0 (so impacts are constant within each school), the blocked design reduces to the basic PN-RCT design with random IC effects.

Table 8 replicates table 7 for the blocked PN-RCT design. We assume for these calculations that $p_S=.5$ and $\omega=.5$, where other parameter values are the same as for table 7. We find that required samples are considerably smaller for this design than for the C-RCT design. For example, if $MDE=.30$, $m_H=5$, $J=2$, and $R^2=.5$, the blocked PN-RCT design requires only 37 schools compared to 62 schools for the C-RCT design. These precision gains occur because random assignment of students is conducted within schools rather than at the school level. In effect, we have a mini-experiment within each school. Of course, random assignment within schools has the potential problem that intervention effects could "spill over" from treatment to control students, which could lead to contaminated impact estimates (see, for example, Rhoads 2011).

Table 8. Total sample size calculations for schools for the blocked PN-RCT design with random school and IC effects for treatment and control groups of equal size

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Average IC sample size ( $J$ ) | 0 | .25 | .50 | .75 |
| **MDE Target = .10; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 593 | 445 | 296 | 148 |
| 2 | 659 | 494 | 329 | 165 |
| **MDE Target = .10; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 192 | 144 | 96 | 48 |
| 2 | 214 | 160 | 107 | 53 |
| 5 | 240 | 180 | 120 | 60 |
| 10 | 284 | 213 | 142 | 71 |
| **MDE Target = .10; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 126 | 94 | 63 | 31 |
| 2 | 140 | 105 | 70 | 35 |
| 5 | 153 | 114 | 76 | 38 |
| 10 | 174 | 131 | 87 | 44 |
| 20 | 218 | 164 | 109 | 55 |
| **MDE Target = .20; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 148 | 111 | 74 | 37 |
| 2 | 165 | 123 | 82 | 41 |
| **MDE Target = .20; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 48 | 36 | 24 | 12 |
| 2 | 53 | 40 | 27 | 13 |
| 5 | 60 | 45 | 30 | 15 |
| 10 | 71 | 53 | 35 | 18 |
| **MDE Target = .20; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 31 | 24 | 16 | 8 |
| 2 | 35 | 26 | 17 | 9 |
| 5 | 38 | 29 | 19 | 10 |
| 10 | 44 | 33 | 22 | 11 |
| 20 | 55 | 41 | 27 | 14 |
| **MDE Target = .30; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 66 | 49 | 33 | 16 |
| 2 | 73 | 55 | 37 | 18 |
| **MDE Target = .30; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 21 | 16 | 11 | 5 |
| 2 | 24 | 18 | 12 | 6 |
| 5 | 27 | 20 | 13 | 7 |
| 10 | 32 | 24 | 16 | 8 |
| **MDE Target = .30; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T} = 0$ | 14 | 10 | 7 | 3 |
| 2 | 16 | 12 | 8 | 4 |
| 5 | 17 | 13 | 8 | 4 |
| 10 | 19 | 15 | 10 | 5 |
| 20 | 24 | 18 | 12 | 6 |

**Table 8. Total sample size calculations for schools for the blocked PN-RCT design with random school and IC effects for treatment and control groups of equal size (Continued)**

| | Regression $R^2$ value from model covariates | | | |
|---|---|---|---|---|
| Average IC sample size ( $J$ ) | 0 | .25 | .50 | .75 |
| **MDE Target = .40; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T}$ = 0 | 37 | 28 | 19 | 9 |
| 2 | 41 | 31 | 21 | 10 |
| **MDE Target = .40; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T}$ = 0 | 12 | 9 | 6 | 3 |
| 2 | 13 | 10 | 7 | 3 |
| 5 | 15 | 11 | 7 | 4 |
| 10 | 18 | 13 | 9 | 4 |
| **MDE Target = .40; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T}$ = 0 | 8 | 6 | 4 | 2 |
| 2 | 9 | 7 | 4 | 2 |
| 5 | 10 | 7 | 5 | 2 |
| 10 | 11 | 8 | 5 | 3 |
| 20 | 14 | 10 | 7 | 3 |
| **MDE Target = .50; $m_H$ = 5*** | | | | |
| 1; $\rho_{\theta T}$ = 0 | 24 | 18 | 12 | 6 |
| 2 | 26 | 20 | 13 | 7 |
| **MDE Target = .50; $m_H$ = 20*** | | | | |
| 1; $\rho_{\theta T}$ = 0 | 8 | 6 | 4 | 2 |
| 2 | 9 | 6 | 4 | 2 |
| 5 | 10 | 7 | 5 | 2 |
| 10 | 11 | 9 | 6 | 3 |
| **MDE Target = .50; $m_H$ = 40*** | | | | |
| 1; $\rho_{\theta T}$ = 0 | 5 | 4 | 3 | 1 |
| 2 | 6 | 4 | 3 | 1 |
| 5 | 6 | 5 | 3 | 2 |
| 10 | 7 | 5 | 3 | 2 |
| 20 | 9 | 7 | 4 | 2 |

**NOTES:** All calculations were conducted using equations (4.9) and (4.14) assuming a 5 percent significance level, two-tailed test at 80 percent power, $\rho_{\xi T} = .15$, $\rho_{\theta T} = .10$ (except where otherwise noted), and equal treatment and control group school sample sizes (that is, $p_H = .5$). For simplicity, all calculations assume $Factor(.) = 2.802$ regardless of the degrees of freedom. The figures in the table show total sample sizes for *schools* (split evenly between the treatment and control groups) that are required to achieve the indicted precision targets measured in effect size units. The figures are shown for various MDE targets, number of students per school $(m_H)$, the average IC sample size $J$, school-level ICCs ( $\rho_{\xi T}$ ), IC-level ICCs $(\rho_{\theta T})$, and regression $R^2$ values.

* Sample sizes for *students* can be obtained by multiplying sample sizes for *schools* by the number of students per school ( $m_H$ ).

Finally, the analysis above can be generalized to the blocked C-RCT design, where schools or classrooms are randomly assigned within higher units, such as school districts, although the notation becomes onerous. For example, in specifications where both district and IC effects are treated as random, the variance of an impact estimate is

$$Var\left(Impact\right)=\frac{\sigma_{\varphi}^{2}}{D}+\left[\frac{\sigma_{\xi}^{2}}{D\times H_{D}\times p_{D}}+\frac{\sigma_{\theta}^{2}}{D\times H_{D}\times p_{D}\times I_{TH}}+\frac{\sigma_{\varepsilon}^{2}}{n_{T}}\right]+\left[\frac{\sigma_{\xi}^{2}}{D\times H_{D}\times(1-p_{D})}+\frac{\sigma_{\varepsilon}^{2}}{n_{C}}\right], \tag{4.17}$$

where $\sigma_{\varphi}^{2}$ is the variance of impacts across districts, $D$ is the number of districts in the sample, $H_D$ is the average number of schools per district, $p_D$ is the average proportion of schools per district that is assigned to the treatment group, and other parameters are defined as above. The design effect for this design can be calculated by dividing this expression by the variance of the reference design in equation (4.7) to yield

$$deff=1+\frac{\lambda_{\pi T}\left(\omega_{D}m_{D}p_{D}(1-p_{D})-1\right)+\lambda_{\xi T}\left(m_{H}-1\right)+\lambda_{\theta T}\left(1-p_{D}\right)(J-1)}{(1-\lambda_{\theta T}p_{D})}, \tag{4.18}$$

where $m_D$ is the average number of students per district; $\lambda_q$ are ICCs of the form $\lambda_{q}=\sigma_{q}^{2}/\sigma_{\lambda}^{2}$ where $\sigma_{\lambda}^{2}=(\sigma_{D}^{2}+\sigma_{\xi}^{2}+\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2})$ and $q\epsilon(D,\xi,\theta,\varepsilon)$; $\sigma_D^2$ is the variance of the random district effects; $\omega_{D}=\sigma_{\pi}^{2}/\ \sigma_{D}^{2}$; and other parameters are defined as above.

## 4.4 Summary of Statistical Power Considerations

The main finding from this chapter is that if researchers aim to treat IC effects as random, student and school sample sizes to achieve precise impact estimates must be somewhat larger under PN-RCT and related designs than under traditional RCT designs without ICs. The main reason this occurs is that IC effects increase the variances of the impact estimates for the treatment group. For similar reasons, the sample size formulas are more complex for PN-RCT designs than for traditional RCT designs, although similar methods can be used to obtain formulas for each type of design.

Another important take away message is that required sample sizes are much smaller if random assignment is conducted within schools rather than at the school level. Thus, if potential study

contamination due to spillover effects is not a concern, within-school random assignment designs should be considered. This same finding also holds for traditional RCTs.

Finally, to apply the formulas developed in the chapter, it is necessary to input plausible ICC values at the IC level. An important area for future research is to obtain empirical values for these ICCs in multiple settings, so education researchers can use appropriate values when planning for their PN-RCTs.

# References


Baldwin S.A., Murray D.M., and Shadish W.R. (2005). Empirically Supported Treatments or Type 1 Errors? Problems With the Analysis of Data From Group-Administered Treatments. *Journal of Consulting and Clinical Psychology, 73*: 924-935.

Bates, D.M., Maechler, M., and Bolker, B. (2012). Lme4: Linear Mixed-Effects Models Using S4 Classes. Version 0.999999-0. Contributed R package. Retrieved February 26, 2013, from http://cran.r-project.org/web/packages/lme4/index.html

Bauer, D.J., Sterba, S.K., and Hallfors, D.D. (2008). Evaluating Group-Based Interventions When Control Participants Are Ungrouped. *Multivariate Behavioral Research, 43*(2): 210-236.

Bloom, H.S., Richburg-Hayes, L., and Black, A.R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*(1): 30-59.

Brogan, D.R., and Kutner, M.H. (1980). Comparative Analyses of Pretest-Posttest Research Designs. *The American Statistician, 34*: 229-232.

Cameron, A.C., Gelbach, J.B., and Miller, D.L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, *29*(2): 238-249.

Campbell, M.K., Elbourne, D.R., and Altman, D.G. (2004). CONSORT Statement: Extension to Cluster Randomised Trials. *British Medical Journal, 328*: 702-708.

Chaplin, D., and Capizzano, J. (2006). *Impacts of a Summer Learning Program: A Random Assignment Study of Building Educated Leaders for Life (BELL): Final Report.* Washington DC: Urban Institute.

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cornfield, J. (1978). Randomization by Group: A Formal Analysis. *American Journal of Epidemiology, 108:* 100-102.

Crits-Christoph, P., and Mintz, J. (1991). Implications of Therapist Effects for the Design and Analysis of Comparative Studies of Psychotherapies. *Journal of Consulting and Clinical Psychology, 59:* 20-26.

Crits-Christoph, P., Tu, X., and Gallop, R. (2003). Therapists as Fixed Versus Random Effects—Some Statistical and Conceptual Issues: A Comment on Siemer and Joormann. (2003). *Psychological Methods, 8:* 518-523.

Demidenko, E. (2004). *Mixed Models.* New York, NY: Wiley.

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

Donner, A., and Klar, N. (2004). Pitfalls of and Controversies in Cluster Randomization Trials. *American Journal of Public Health*, *94:* 416-422.

Dudewicz, E.J., Ma, Y., Mai, E., and Su. H. (2007). Exact Solutions to the Behrens-Fisher Problem: Asymptotically Optimal and Finite Sample Choice Among. *Journal of Statistical Planning and Inference, 137:* 1584-1605.

Dugard, P., and Todman, J. (1995). Analysis of Pre-Test-Post-Test Control Group Designs in Educational Research. *Educational Psychology, 15*: 181-198.

Dyson, N.I., Jordan, N.C., and Glutting, J. (2013). A Number Sense Intervention for Urban Kindergarteners At-Risk for Mathematics Difficulties. *Journal of Learning Disabilities*, *46*(2): 166-181.

Fai, A.H.T., and Cornelius, P.L. (1996). Approximate *F*-tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-Plot Experiments. *Journal of Statistical Computation and Simulation, 54:* 363-378.

Fielding, A., and Goldstein, H. (2006). *Cross-Classified and Multiple Membership Structures in Multi-Level Models: An Introduction and Review* (No. 791). Birmingham, UK.: University of Birmingham. Department of Education and Skills.

Gates, C.E. (1995). What Really Is Experimental Error in Block Designs? *American Statistician*, *49:* 362-363.

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., Grider, M., and Britton, E. (2010). *Impacts of Comprehensive Teacher Induction: Final Results From a Randomized Controlled Study* (NCEE 2010–4027). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Goldstein, H. (1995). *Multilevel Statistical Models* (2nd ed.). New York, NY: John Wiley.

Hedges, L., and Hedberg, E. (2007). Intraclass Correlation Values for Planning Group Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, *29:* 60-87.

Hedges, L., and Rhoads, C. (2009). *Statistical Power Analysis in Education Research* (NCSER 2010-3006). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research.

Hoover, D.R. (2002). Clinical Trials of Behavioural Interventions With Heterogeneous Teaching Subgroup Effects. *Stat Med.*, *21:* 1351-1363.

James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., and Pistorino, C. (2005). *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program: Final Report.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Jenney, B., and Lohr, S. (2009). Experimental Designs for Multiple-Level Responses, With Application to a Large-Scale Educational Intervention. *Annals of Applied Statistics*, *3*(2): 691-709.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York, NY: Springer.

Karl, A., Yang, Y., and Lohr, S. (2013). Efficient Maximum Likelihood Estimation of Multiple Membership Linear Mixed Models, With an Application to Educational Value-Added Assessments. *Computational Statistics and Data Analysis*, *59:* 13-27.

Kiernan, K., Tao, J., and Gibbs, P. (2012). Tips and Strategies for Mixed Modeling With SAS/STAT® Procedures. *Proceedings of the SAS® Global Forum 2012 Conference*, Cary, NC: SAS Institute Inc. Retrieved 3/2/2013 from http://support.sas.com/resources/papers/proceedings12/332-2012.pdf

Laird, N. (1983). Further Comparative Analyses of Pretest-Posttest Research Designs. *The American Statistician,* 37: 329-330.

Lee, K. J., and Thompson, S.G. (2005). The Use of Random Effects Models to Allow for Clustering in Individually Randomized Trials. *Clinical Trials*, *2:* 163-173.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K S., and Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms*. (NCSER 2013-3000). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research.

Lockwood, J.R., Weiss, M., and McCaffrey, D.F. (2013). Estimating the Standard Error of the Impact Estimate in Individually Randomized Trials, with Clustering. Paper presented at the Fall 2013 Society for Research on Educational Effectiveness (SREE) Conference, Washington DC, https://www.sree.org/download/files/1385572216t4_conf_1044.pdf

Lohr, S. (1995). Hasse Diagrams in Statistical Consulting and Teaching, *The American Statistician*, *49:* 376-381.

Lohr, S., and Divan, M. (1997). Comparison of Confidence Intervals for Variance Components With Unbalanced Data. *Journal of Statistical Computation and Simulation*, *58:* 83-97.

Martindale, C. (1978). The Therapist-As-Fixed-Effect Fallacy in Psychotherapy Research. *Journal of Consulting and Clinical Psychology, 46:* 1526-1530.

McLean, R.A., Sanders, W.L., and Stroup, W.W. (1991). A Unified Approach to Mixed Linear Models. *The American Statistician*, *45:* 54-64.

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

National Forum on Education Statistics (2010). *Forum Guide to Data Ethics (NFES 2010-801).* U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2010/2010801.pdf.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org.

Raudenbush, S. (1993). A Crossed Random Effects Model for Unbalanced Data With Applications in Cross-Sectional and Longitudinal Research. *Journal of Educational Statistics, 18*(4): 321-349.

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

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

Raudenbush, S., and Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W., Bryk, A S., and Congdon Jr., R.T. (2011). *HLM 7.0 for Windows*. Skokie, IL: Scientific Software International.

Rhoads, C. (2011) The Implications of Contamination for Experimental Design in Education Research. *Journal of Educational and Behavioral Statistics,*36(1), 76-104.

Roberts, C., and Roberts, S.A. (2005). Design and Analysis of Clinical Trials With Clustering Effects Due to Treatment. *Clinical Trials*, *2:* 152-162.

Roberts, J., Williams, K., Carter, M., Evans, D., Parmenter, T., Silove, N., Clark, T., & Warren, A. (2011). A randomized controlled trial of two early intervention programs for young children with autism: Centre-based with parent program and home-based. *Research in Autism Spectrum Disorders, 5,* 1553-1566.

Rolfhus, E., Gersten, R., Clarke, B., Decker, L., Wilkins, C., and Dimino, J. (2012). *An Evaluation of Number Rockets: A Tier 2 Intervention for Grade 1 Students At Risk for Difficulties in Mathematics* (NCEE 2012-4007). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_20124007.pdf.

Sanders, E. (2011). *Multilevel Analysis Methods for Partially Nested Cluster Randomized Trials* (Doctoral dissertation). Ann Arbor, MI: ProQuest LLC.

SAS Institute, Inc. (2011). *SAS/STAT® 9.3 User's Guide: The MIXED Procedure (Chapter).* Cary, NC: Author.

Satterthwaite, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2: 110-114.

Schabenberger, O. (2004). Mixed Model Influence Diagnostics. *Proceedings of the Twenty-Ninth Annual SAS® Users Group International Conference.* Cary, NC: SAS Institute Inc. Retrieved March 2, 2013, from http://www2.sas.com/proceedings/sugi29/189-29.pdf

Schochet, P.Z. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*(1): 62-87.

Schochet, P. Z. (2009) "Statistical Power for Regression Discontinuity Designs in Education Evaluations." *Journal of Educational and Behavioral Statistics, 34(2)*, 2009, 238-266.

Schochet, P. Z. (2011) "Do Typical RCTs of Education Interventions Have Sufficient Statistical Power for Linking Impacts on Teacher and Student Outcomes?" *Journal of Educational and Behavioral Statistics, 36(4)*, pp. 441-471.

Self, S.G., and Liang, K.Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, *82:* 605-610.

Serlin, R.C., Wampold, B.E., and Levin, J.R. (2003). Should Providers of Treatment Be Regarded as a Random Factor? If It Ain't Broke, Don't "Fix" It: A Comment on Siemer and Joormann (2003). *Psychological Methods*, *8:* 524-534.

Siemer, M., and Joormann, J. (2003). Power and Measures of Effect Size in Analysis of Variance With Fixed Versus Random Nested Factors. *Psychological Methods, 8:* 497-517.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC.

Walsh, J.E. (1947). Concerning the Effect of Intraclass Correlation on Certain Significance Tests. *The Annals of Mathematical Statistics*, *18:* 88-96.

Wampold, B.E., and Serlin, R.C. (2000). The Consequence of Ignoring a Nested Factor on Measures of Effect Size in Analysis of Variance. *Psychological Methods, 5:* 425-433.

**This page left blank for double-sided copying.**

# Appendix A

# Mixed Model Theory for PN-RCTs

**This page left blank for double-sided copying.**

# Appendix A
# Mixed Model Theory for PN-RCTs

All of the models discussed in chapter 3 and 4 may be written in the form of a general mixed model. In this Appendix, we express each model in standard mixed model form, using the notation in the "Mixed Models Theory" section of the PROC MIXED documentation in SAS Institute Inc. (2011). We use standard convention that a boldface letter denotes a matrix or vector, and standard matrix notation for transpose $(')$ and inverse $(^{-1})$ of a matrix. Let $\boldsymbol{I}_r$ denote the $r \times r$ identity matrix, let $0_{rc}$ denote the $r \times c$ matrix with each entry equal to zero, and let $\mathbf{1_{rc}}$ denote the $r \times c$ matrix with each entry equal to one.

The individual models have different numbers of subscripts describing the student response. To achieve a uniform notation, we refer to each student using the subscript $l$ rather than the subscripts denoting the full hierarchical structure: the hierarchical structure for each model will be defined by the random effects and the covariances for each model separately. The outcome for student $l$ may be written as

$$y_l = \beta_0 + \beta_1 T_l + \boldsymbol{z}_l' \boldsymbol{\gamma} + \varepsilon_l \tag{A.1}$$

where $T_l = 1$ if the student is in the treatment group and $T_l = 0$ if the student is in the control group. The vector $\boldsymbol{\gamma}$ contains all of the random effects, $\boldsymbol{z}_l$ describes which random effects are associated with the response of student $l$, and $\varepsilon_l$ is an error term associated with student $l$'s individual response, assumed to be independently normally distributed with mean 0 and variance

$$Var(\varepsilon_l) = \begin{cases} \sigma^2_{\varepsilon C} \text{ if } T_l = 0 \\ \sigma^2_{\varepsilon T} \text{ if } T_l = 1 \end{cases}$$

The parameter $\beta_1$ is the impact of the treatment on student outcomes (ATE). For all designs considered, the generalized least squares estimator $\hat{\beta}_1$ of $\beta_1$ will be derived, along with its theoretical variance.

This model is a special case of the general mixed model (see Demidenko 2004; SAS Institute 2011),

$$\boldsymbol{y} = \boldsymbol{X\beta} + [\boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}] \tag{A.2}$$

where $\boldsymbol{y} = (y_1, y_2 \ldots, y_n)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the vector of fixed effects, $\boldsymbol{\gamma}$ is a $K$-vector of random effects, $\boldsymbol{X}$ is the $n \times 2$ matrix whose $l$th row is $(1, T_l)$, $\boldsymbol{Z}$ is the $n \times K$ matrix whose $l$th row is $\boldsymbol{z}_l'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ is the vector of error terms.[26] We assume that $\boldsymbol{\gamma} \sim N(0, \boldsymbol{G})$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{R})$ are independent so the $n \times n$ variance-covariance matrix of the vector of responses is

$$\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}. \tag{A.3}$$

The fixed effects $\boldsymbol{\beta}$ are estimated by the empirical best linear unbiased estimator

$$\boldsymbol{\beta} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \tag{A.4}$$

where $\hat{\boldsymbol{V}}$ is a consistent estimator of the theoretical variance matrix $\boldsymbol{V}$. Typically, maximum likelihood (ML) or restricted maximum likelihood (REML) methods are used to obtain estimates of the components of $\boldsymbol{V}$ that are substituted into $\hat{\boldsymbol{V}}$.

The variance of the fixed effects estimates is estimated by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}.$$

This estimator, which is typically used in practice, may have a slight downward bias because it does not account for the variability in estimating the components of $V$. This bias is small, however, for large or well balanced data sets (SAS Institute, 2011, p. 4803).

The structures of the random effects vector $\boldsymbol{\gamma}$, its associated descriptive vector $\boldsymbol{z}_i$, and the matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are the only features that differ among the designs considered. We now express each of the models in the form of Equation (A.2) and derive the properties of the estimator of the ATE, $\hat{\beta}_1$, for each model. We shall work through the model for the basic PN-RCT design in detail, then provide the variance structure and properties of the ATE estimator for the other models.

---

[26] The model in equation (A.1) includes only the fixed effects for the intercept and the treatment effect. It is easily extended to include additional covariates by appending fixed effects terms. In that case, $\boldsymbol{\beta}$ will be a vector of length $p$ (the total number of fixed effects parameters), and $\boldsymbol{X}$ will be an $n \times p$ matrix. The variance structures presented in the Appendix will be unchanged, and the empirical best linear unbiased estimator for $\boldsymbol{\beta}$ will be given by Equation (A.4), using the larger $\boldsymbol{X}$ matrix.

## A.1 Equation (3.3) for Basic PN-RCT Design

For Equation (3.3), order the data so the first $n_C$ observations are in the control group, followed by the treatment group observations in ICs 1 through $I_T$, where IC $i$ has $J_i$ students. The sample consists of a total of $n = n_C + n_T$ observations. The control group has no intervention clusters, and each IC has its own random effect in the treatment group, so we set $\boldsymbol{\gamma} = (0, \theta_1, \theta_2 \ldots, \theta_{I_T})'$. To capture the correct IC effect for each student, set $\boldsymbol{z}_l = (1,0,0,\ldots,0)'$ if student $l$ is in the control group and $\boldsymbol{z}_l = (0,0,\ldots,0,1,0,\ldots,0)'$, with a 1 in position $(i+1)$ and zeroes elsewhere, if student $l$ is in IC $i$ in the treatment group. Since it is assumed that $\theta_i \sim N(0, \sigma_\theta^2)$ are independent for $i = 1$ to $I_T$, $\boldsymbol{G}$ is the $(I_T + 1) \times (I_T + 1)$ diagonal matrix:

$$\boldsymbol{G} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} 0 & 0_{1,I_T} \\ 0_{I_T,1} & \sigma_\theta^2 \boldsymbol{I}_{I_T} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & \sigma_\theta^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_\theta^2 \end{bmatrix}.$$

The $\boldsymbol{G}$ matrix in this formulation is singular because the first diagonal entry is 0. It is thus non-negative definite but not positive definite.

The covariance matrix for the student-level error terms is also diagonal, with diagonal entry $\sigma_{\varepsilon C}^2$ for each student in the control group and diagonal entry $\sigma_{\varepsilon T}^2$ for each student in the treatment group. Thus, $\boldsymbol{R}$ is an $n \times n$ matrix with the form:

$$\boldsymbol{R} = Var(\boldsymbol{\varepsilon}) = diag\left(\sigma_{\varepsilon C}^2, \ldots, \sigma_{\varepsilon C}^2, \sigma_{\varepsilon T}^2, \ldots, \sigma_{\varepsilon T}^2\right) = \begin{bmatrix} \sigma_{\varepsilon C}^2 \boldsymbol{I}_{n_C} & \boldsymbol{0}_{n_C, n_T} \\ \boldsymbol{0}_{n_T, n_C} & \sigma_{\varepsilon T}^2 \boldsymbol{I}_{n_T} \end{bmatrix}. \quad \text{(A.5)}$$

This structure leads to a block diagonal form for the $n \times n$ covariance matrix, $\boldsymbol{V}$:

$$\boldsymbol{V} = Var(\boldsymbol{y}) = \boldsymbol{ZGZ}' + \boldsymbol{R} = \begin{bmatrix} \boldsymbol{V}_C & 0 & \ldots & 0 \\ 0 & \boldsymbol{V}_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{V}_{I_T} \end{bmatrix}.$$

Because students in the control group are independent, $\boldsymbol{V}_C = \sigma^2_{\varepsilon C}\boldsymbol{I}_{n_C}$. The matrix $\boldsymbol{V}_i$ for each IC in the treatment group reflects the clustering induced by the ICs: $\boldsymbol{V}_i$ is the $J_i \times J_i$ matrix with diagonal entries $\sigma^2_{\varepsilon T} + \sigma^2_\theta$ and off-diagonal entries $\sigma^2_\theta$, written as

$$\boldsymbol{V}_i = \sigma^2_{\varepsilon T}\boldsymbol{I}_{J_i} + \sigma^2_\theta 1_{J_i,J_i} = \begin{bmatrix} \sigma^2_{\varepsilon T}+\sigma^2_\theta & \sigma^2_\theta & \cdots & \sigma^2_\theta \\ \sigma^2_\theta & \sigma^2_{\varepsilon T}+\sigma^2_\theta & \cdots & \sigma^2_\theta \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_\theta & \sigma^2_\theta & \cdots & \sigma^2_{\varepsilon T}+\sigma^2_\theta \end{bmatrix}$$

This variance structure implies that the covariance among students in the same IC is $\sigma^2_\theta$ and the intraclass correlation of students in the same IC is $\rho_\theta = \sigma^2_\theta / \ (\sigma^2_{\varepsilon T} + \sigma^2_\theta)$. Note that the same variance structure is assumed to hold in each IC.

Note that $\boldsymbol{V}_C^{-1} = \dfrac{1}{\sigma^2_{\varepsilon C}}\boldsymbol{I}_{n_C}$ and

$$\boldsymbol{V}_i^{-1} = \frac{1}{\sigma^2_{\varepsilon T}}\left(\boldsymbol{I}_{J_i} - \frac{\sigma^2_\theta}{\sigma^2_{\varepsilon T} + J_i\sigma^2_\theta}\mathbf{1}_{\mathbf{J_i},\mathbf{J_i}}\right).$$

Using Equation (A.4), then,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \overline{y}_C \\ \sum_{i=1}^{I_T}\hat{\omega}_i\overline{y}_i - \overline{y}_C \end{bmatrix},$$

where $\overline{y}_C$ is the mean of the control group students, $\overline{y}_i$ is the mean of the treatment group students in IC $i$, and

$$\hat{\omega}_i = \left(\frac{1}{(\hat{\sigma}^2_{\varepsilon T}/J_i) + \hat{\sigma}^2_\theta}\right) \Big/ \left(\sum_{k=1}^{I_T}\frac{1}{(\hat{\sigma}^2_{\varepsilon T}/J_k) + \hat{\sigma}^2_\theta}\right)$$

is the weight accorded to IC $i$ when estimating the mean of the treatment group. This results in precision weighting, where larger ICs get a larger weight in the calculation of the overall treatment group mean than smaller ICs (although this effect is dampened somewhat by the common $\hat{\sigma}^2_\theta$ term that is not deflated by the IC size). Note that if all the ICs have the same size, i.e.,

$J_i = J_2 = \ldots = J_{I_T} = J$, then $\hat{\omega}_i = 1/I_T$ and $\hat{\beta}_1 = \bar{y}_T - \bar{y}_C$, where $\bar{y}_T$ is the simple average of the IC means. Also note that if the variance due to ICs is zero, then $\hat{\omega}_i = J_i / \sum_{k=1}^{I_T} J_k$ so the estimated mean of the treatment group is a weighted average of the individual IC means, where the weights are proportional to IC size.

We now derive the variance of the estimator. To facilitate the theory, we derive the variance of the best linear unbiased estimator,

$$\tilde{\boldsymbol{\beta}} = \left( \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y},$$

which is the estimator of the same form as $\hat{\boldsymbol{\beta}}$ but assumes that the variance components are known. For the model in Equation (3.3),

$$Var(\tilde{\boldsymbol{\beta}}) = \left( \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} = \begin{bmatrix} \dfrac{\sigma^2_{\varepsilon C}}{n_C} & -\dfrac{\sigma^2_{\varepsilon C}}{n_C} \\ -\dfrac{\sigma^2_{\varepsilon C}}{n_C} & \dfrac{\sigma^2_{\varepsilon C}}{n_C} + 1/\left( \displaystyle\sum_{k=1}^{I_T} \frac{1}{(\sigma^2_{\varepsilon T}/J_k) + \sigma^2_\theta} \right) \end{bmatrix}.$$

Thus, the variance of the ATE is

$$Var(\tilde{\beta}_1) = \frac{\sigma^2_{\varepsilon C}}{n_C} + 1/\left( \sum_{k=1}^{I_T} \frac{1}{(\sigma^2_{\varepsilon T}/J_k) + \sigma^2_\theta} \right). \qquad \text{(A.6)}$$

If all the ICs have the same size, i.e., $J_1 = J_2 = \ldots = J_{I_T} = J$, then

$$Var(\tilde{\beta}_1) = \frac{\sigma^2_{\varepsilon C}}{n_C} + \frac{\sigma^2_{\varepsilon T}}{n_T} + \frac{\sigma^2_\theta}{I_T}$$

which is the expression given in section 3.2 and used in chapter 4. Note that when $\sigma^2_\theta > 0$, the size of the variance of the estimator depends on the number of ICs. If $I_T$ is small relative to $n_T$ and $n_C$, then the variance will be dominated by the $\dfrac{\sigma^2_\theta}{I_T}$ term no matter how many students are in the control group or how many students are in each IC. Results in Demidenko (2004, chapter 3) can be used to

show that $Var(\tilde{\beta}_1) \approx Var(\tilde{\beta}_1)$ for either ML or REML estimators of the variance parameters when $I_T$ is large.

The basic model structure for the basic PN-RCT can be modified for the situation described in section 4.2.1 in which a student may attend multiple ICs. Instead of defining $\boldsymbol{z}_l = (0,0,\ldots,0,1,0,\ldots,0)'$ for a treatment group student, with a 1 corresponding to the IC membership, allow the entries of $\boldsymbol{z}_l$ to correspond to the fraction of time that student $l$ spent in each IC. Thus, a student who spent half of the time in IC 1 and half of the time in IC 2 would have $\boldsymbol{z}_l = (0,1/2\ ,1/2\ ,0,\ldots,0)'$. Note that this change in $\boldsymbol{z}_l$ also changes the correlation structure because treatment group students are no longer nested in ICs. Equation (A.3) must be used to find the covariance matrix for the observations.

## A.2 Equations (3.5) and (3.6) for Blocked PN-RCT Design

For the blocked PN-RCT design, order the observations by school, with the schools labeled as 1 to $H$. Within schools, order the observations by IC, with the control students followed by the students in each of the ICs in that school. This model can be expressed in the general form of the models in (A.1) and (A.2) by defining the vectors $\boldsymbol{z}_l$ and $\boldsymbol{\gamma}$. The covariance structure becomes more complex because each school now contains both control and treatment group students.

The random effects in the vector $\boldsymbol{\gamma}$ now consist of the school-level effects $\xi_h$ and the IC-level effects $\theta_{hi}$ within each of the $H$ schools, with respective variances $\sigma^2_\xi$ and $\sigma^2_\theta$. We thus define

$$\boldsymbol{\gamma} = (\xi_1,\ldots,\xi_H,0,\theta_{11},\theta_{12}\ldots,\theta_{1,I_{T1}},\ldots,\theta_{H,1},\theta_{H,2}\ldots,\theta_{H,I_{TH}})'$$

which lists the $H$ school effects, followed by 0 to represent the IC effect for the control students in each school (IC 0 within each school), and then followed by a separate set of IC effects for each IC in the $H$ schools. For all students, set $\boldsymbol{h}_l = (0,0,\ldots,0,1,0,\ldots,0)'$ to be the $H$-vector with a 1 in position $h$ if student $l$ is in school $h$. To account for the IC effects in the treatment (but not control) group, let $\boldsymbol{c}_l = (1,0,0,\ldots,0)'$ if student $l$ is in the control group of school $h$, and $\boldsymbol{c}_l = (0,0,\ldots,0,1,0,\ldots,0)'$, with a 1 in position $(i+1)$ and zeroes elsewhere, if student $l$ is in IC $i$ within school $h$. Then $\boldsymbol{z}_l = (\boldsymbol{h}_l', \boldsymbol{c}_l')'$. Since it is assumed that all variance components are independent, for this model $\boldsymbol{G}$ is the $\left(H+1+\sum_{h=1}^{H} I_{Th}\right) \times \left(H+1+\sum_{h=1}^{H} I_{Th}\right)$ diagonal matrix:

$$\boldsymbol{G} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} \sigma_\xi^2 \boldsymbol{I}_H & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\theta^2 \mathbf{I}_{\sum_{h=1}^{H} I_{TH}} \end{bmatrix}.$$

The student-level covariance matrix, $\boldsymbol{R}$, is, as before, the $n \times n$ matrix given in Equation (A.5).

The $n \times n$ covariance matrix $\boldsymbol{V}$ is again block diagonal:

$$\boldsymbol{V} = Var(\boldsymbol{y}) = \boldsymbol{ZGZ'} + \boldsymbol{R} = \begin{bmatrix} \boldsymbol{V}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{V}_H \end{bmatrix}$$

where the submatrices $\boldsymbol{V}_1$ through $\boldsymbol{V}_H$ give the covariance structure within each of the $H$ schools. The blocked PN-RCT design consists of $H$ independent replicates of the basic PN-RCT design, so each submatrix $\boldsymbol{V}_h$ has the form of the $\boldsymbol{V}$ matrix from the basic PN-RCT, with the addition of a school-level random effect in each entry of the submatrix:

$$\boldsymbol{V}_h = \begin{bmatrix} \boldsymbol{V}_{hC} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_{h1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{V}_{hI_{Th}} \end{bmatrix} + \sigma_\xi^2 \mathbf{1}_{m_h, m_h}$$

where $m_h$ is the total number of students in school $h$. In this matrix, $\boldsymbol{V}_{hC} = \sigma_{\varepsilon C}^2 \boldsymbol{I}$ and $\boldsymbol{V}_{hi} = \sigma_{\varepsilon T}^2 \boldsymbol{I} + \sigma_\theta^2 \mathbf{1}_{J_{hi}, J_{hi}}$, where $J_{hi}$ is the number of students in IC $i$ of school $h$.

Note that the block diagonal structure of the matrix $\boldsymbol{V}$ implies that students from different schools are independent. All students in the *same* school, however, are positively correlated. The covariance of two control students, or the covariance of two students in different ICs, is $\sigma_\xi^2$. The covariance of two treatment students in the same IC is $\sigma_\xi^2 + \sigma_\theta^2$.

The variance structure for Equation (3.6) is similar, with the addition of the extra covariance terms in $\boldsymbol{G}$ and $\boldsymbol{V}$. Because we allow the school-level random effects, **G** is no longer a diagonal matrix: we form $\boldsymbol{z}_l = (\boldsymbol{h}_l', \boldsymbol{c}_l')'$ as above but now modify $\boldsymbol{h}_l$ to be the $2H$ -vector with a 1 in position $2h-1$

if student $l$ is in school $h$ and in the control group, and 1's in positions $2h-1$ and $2h$ if student $l$ is in school $h$ and in the treatment group. Then,

$$\boldsymbol{G} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{G}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\theta^2 \mathbf{I}_{\sum_{h=1}^{H} I_{TH}} \end{bmatrix}.$$

where $\boldsymbol{G}_1$ is block diagonal consisting of $H$ blocks $\boldsymbol{G}_2$:

$$\boldsymbol{G}_2 = \begin{bmatrix} \sigma_\xi^2 & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma_\eta^2 \end{bmatrix}. \tag{A.7}$$

This structure for **G** results in the following covariance structure for the students in school $h$:

$$\boldsymbol{V}_h = \begin{bmatrix} \boldsymbol{V}_{hC} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_{h1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{V}_{hI_{Th}} \end{bmatrix} + \sigma_\xi^2 \mathbf{1}_{m_h, m_h} + \begin{bmatrix} \mathbf{0} & \sigma_{\xi\eta} \mathbf{1}_{m_{hC}, m_{hT}} \\ \sigma_{\xi\eta} \mathbf{1}_{m_{hT}, m_{hC}} & (\sigma_\eta^2 + 2\sigma_{\xi\eta}) \mathbf{1}_{m_{hT}, m_{hT}} \end{bmatrix}$$

For either Equation (3.5) or (3.6), the form of $\hat{\boldsymbol{\beta}}$ in Equation (A.4) is complex for unbalanced data, but simplifies for balanced designs. In the balanced design case, each school has the same number of students ($m_T$ students in the treatment group and $m_C = m_T$ students in the control group), and each IC has the same number $(I_T)$ of students. This implies that the "hat" matrix $\boldsymbol{(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}}$ from Equation (A.4) does not depend on the values of any of the variance components: the first row of $\boldsymbol{(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}}$ contains the value $1/(Hm_C)$ for students in the control group and 0 for students in the treatment group, while the second row of $\boldsymbol{(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}}$ contains the value $-1/(Hm_C)$ for each student in the control group and $1/(Hm_T)$ for each student in the treatment group. Consequently, $\hat{\beta}_0 = \bar{y}_C$ and $\hat{\beta}_1 = \bar{y}_T - \bar{y}_C$. In the balanced case, then, the variance of the ATE is

$$Var(\hat{\beta}_1) = \frac{1}{H^2} \sum_{h=1}^{H} Var(\bar{y}_{Th} - \bar{y}_{Ch}) = \frac{1}{H} \left( \sigma_\eta^2 + \frac{\sigma_\theta^2}{I_T} + \frac{\sigma_{\varepsilon T}^2}{m_T} + \frac{\sigma_{\varepsilon C}^2}{m_C} \right)$$

## A.3 Equation (4.1) for Clustered Design

To express the model in Equation (4.1) in matrix terms, order the observations by school, with the control schools labeled as 1 to $H_C$ and the treatment schools labeled as $H_C+1$ to $H$. Within the treatment schools, order the observations by IC. This model can be expressed in the general form of the models in Equations (A.1) and (A.2) by defining the vectors $\boldsymbol{z}_l$ and $\boldsymbol{\gamma}$.

The random effects in the vector $\boldsymbol{\gamma}$ now consist of the school-level effects $\xi_h$ for the treatment and control groups and the IC-level effects $\theta_{hi}$, with respective variances $\sigma^2_{\xi T}$, $\sigma^2_{\xi C}$ and $\sigma^2_{\theta}$. We thus define

$$\boldsymbol{\gamma} = (\xi_1,\ldots,\xi_{H_C},\xi_{H_C+1},\ldots\xi_H,0,\theta_{(H_C+1),1},\theta_{(H_C+1),2}\ldots,\theta_{(H_C+1),I_{T,(H_C+1)}},\ldots,\theta_{H,1},\theta_{H,2}\ldots,\theta_{H,I_{TH}})'$$

which lists the $H$ school effects, followed by 0 to represent the IC effect for the control schools (inserted to match the SAS formulation), and then followed by a separate set of IC effects for each treatment school. For all students, set $\boldsymbol{h}_l = (0,0,\ldots,0,1,0,\ldots,0)'$ to be the $H$-vector with a 1 in position $h$ if student $l$ is in school $h$. To account for the IC effects in the treatment (but not control) group, let $\boldsymbol{c}_l = (1,0,0,\ldots,0)'$ if student $l$ is in the control group and $\boldsymbol{t}_l = (0,0,\ldots,0,1,0,\ldots,0)'$, with a 1 in position $(i+1)$ and zeroes elsewhere, if student $l$ is in IC $(hi)$ in the treatment group. Then $\boldsymbol{z}_l = (\boldsymbol{h}_l', \boldsymbol{c}_l')'$ if student $l$ is in the control group and $\boldsymbol{z}_l = (\boldsymbol{h}_l', \boldsymbol{t}_l')'$ if student $l$ is in the treatment group. Since it is assumed that all variance components are independent, for this model $\boldsymbol{G}$ is the $\left(H+1+\sum_{h=H_C+1}^{H} I_{Th}\right)\times\left(H+1+\sum_{h=H_C+1}^{H} I_{Th}\right)$ diagonal matrix:

$$\boldsymbol{G} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} \sigma^2_{\xi C}\boldsymbol{I}_{H_C} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2_{\xi T}\boldsymbol{I}_{H_T} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \sigma^2_{\theta}\boldsymbol{I}_{\sum_{h=H_C+1}^{H} I_{Th}} \end{bmatrix}$$

The $\boldsymbol{G}$ matrix in this formulation is singular (as in the basic PN-RCT design) because of the zero diagonal entry. The covariance matrix for the student-level error terms is also diagonal, and has the same form as before, with $\boldsymbol{R}$ the $n\times n$ matrix given in Equation (A.5).

This structure leads to a block diagonal form for the $n\times n$ covariance matrix, $\boldsymbol{V}$:

$$\boldsymbol{V} = Var\left(\boldsymbol{y}\right) = \boldsymbol{ZGZ'} + \boldsymbol{R} = \begin{bmatrix} \boldsymbol{V}_C & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_T \end{bmatrix}.$$

The submatrices $\boldsymbol{V}_C$ and $\boldsymbol{V}_T$ are also block diagonal with diagonal blocks $\boldsymbol{V}_{C1},\ldots,\boldsymbol{V}_{CH_C}$ and $\boldsymbol{V}_{T,H_C+1},\ldots,\boldsymbol{V}_{TH}$ respectively. Each submatrix of $\boldsymbol{V}_C$ has the form:

$$\boldsymbol{V}_{Ch} = \sigma^2_{\varepsilon C}\boldsymbol{I} + \sigma^2_{\xi C}\boldsymbol{1}_{m_{Ch},m_{Ch}} = \begin{bmatrix} \sigma^2_{\varepsilon C}+\sigma^2_{\xi C} & \sigma^2_{\xi C} & \cdots & \sigma^2_{\xi C} \\ \sigma^2_{\xi C} & \sigma^2_{\varepsilon C}+\sigma^2_{\xi C} & \cdots & \sigma^2_{\xi C} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{\xi C} & \sigma^2_{\xi C} & \cdots & \sigma^2_{\varepsilon C}+\sigma^2_{\xi C} \end{bmatrix}$$

for $h=1$ to $H_C$, where $m_{Ch}$ is the number of students in control school $h$. The structure for $\boldsymbol{V}_T$ is more complex because of the additional variance component due to the ICs: for school $h = H_C + 1$ to $H$,

$$\boldsymbol{V}_{Th} = \begin{bmatrix} \sigma^2_{\varepsilon T}\boldsymbol{I} + \sigma^2_{\theta}\boldsymbol{1}_{J_{h1},J_{h1}} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2_{\varepsilon T}\boldsymbol{I} + \sigma^2_{\theta}\boldsymbol{1}_{J_{h2},J_{h2}} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \sigma^2_{\varepsilon T}\boldsymbol{I} + \sigma^2_{\theta}\boldsymbol{1}_{J_{hI_{Th}},J_{hI_{Th}}} \end{bmatrix} + \sigma^2_{\xi T}\boldsymbol{1}_{m_{Th},m_{Th}}$$

The form of $\hat{\boldsymbol{\beta}}$ in Equation (A.4) is complex for unbalanced data, but simplifies for balanced designs. With a balanced design, the "hat" matrix $\left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}$ from Equation (A.4) does not depend on the values of any of the variance components. If each school has the same number of students, and each IC in the treatment group has the same number of students, then $\hat{\beta}_0 = \bar{y}_C$ and $\hat{\beta}_1 = \bar{y}_T - \bar{y}_C$. In the balanced case, the variance of the ATE is

$$Var(\hat{\beta}_1) = Var\left(\bar{y}_C\right) + Var\left(\bar{y}_T\right) = \left(\frac{\sigma^2_{\xi C}}{H_C} + \frac{\sigma^2_{\varepsilon C}}{n_C}\right) + \left(\frac{\sigma^2_{\xi T}}{H_T} + \frac{\sigma^2_{\theta}}{H_T \times I_{H1}} + \frac{\sigma^2_{\varepsilon T}}{n_{CT}}\right)$$

which was used in the power calculations in chapter 4.

## A.4 Equations (4.2) and (4.3) for Design with Randomization of Schools Within Blocks

To express the model in (4.2) in matrix terms, order the observations by district, then by school within district (with control schools preceding treatment schools), IC within treatment schools, and finally by student. We rely on the derivations already done in section A.3 to simplify the presentation of the matrix structure for this design.

As with all designs considered, the covariance matrix $\boldsymbol{V}$ is block diagonal: here, districts are assumed independent of each other. Thus,

$$\boldsymbol{V} = Var(\boldsymbol{y}) = \begin{bmatrix} \boldsymbol{V}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{V}_D \end{bmatrix}.$$

Each submatrix $\boldsymbol{V}_d$, for $d = 1$ to $D$, then gives the covariance structure for students within that district. We can write

$$\boldsymbol{V}_d = \begin{bmatrix} \boldsymbol{V}_{dC} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_{dT} \end{bmatrix} + \sigma_\delta^2 \mathbf{1},$$

where the submatrices $\boldsymbol{V}_{dC}$ and $\boldsymbol{V}_{dT}$ have the same general structure as the submatrices $\boldsymbol{V}_C$ and $\boldsymbol{V}_T$ in section A.2.

The only difference, then, between the variance structure for this design and that for the single district structure in section A.3 is that all of the students within a district have the additional covariance term $\sigma_\delta^2$. This implies that control students in a district are positively correlated with treatment students in the district, which enhances the precision of the ATE because the covariance of the mean treatment student score in a district with the mean control student score is positive.

In a balanced design, the ATE is once again $\hat{\beta}_1 = \bar{y}_T - \bar{y}_C$. For the model in Equation (4.2) with a balanced design,

$$Var(\hat{\beta}_1) = \frac{1}{D^2} \sum_{h=1}^{H} Var\left(\bar{y}_{Td} - \bar{y}_{Cd}\right) = \frac{1}{D}\left(\frac{\sigma_\xi^2}{H_D} + \frac{\sigma_\theta^2}{H_D I_T} + \frac{\sigma_{\varepsilon T}^2}{H_D m_T} + \frac{\sigma_{\varepsilon C}^2}{H_D m_C}\right)$$

where there are $H_D$ schools in each district, $I_T$ ICs in each treatment school, $m_C$ students in each control school, and $m_T = m_C$ students in each treatment school. In this case, the district-to-district variability cancels when taking the difference between the average treatment and control school scores within each district

For the model in Equation (4.3), the leading term of the variance for a balanced design is the additional variability assumed to be due to differential treatment impacts across the districts:

$$Var(\hat{\beta}_1) = \frac{1}{D^2}\sum_{h=1}^{H} Var\left(\bar{y}_{Td} - \bar{y}_{Cd}\right) = \frac{1}{D}\left(\sigma_\varphi^2 + \frac{\sigma_\xi^2}{H_D} + \frac{\sigma_\theta^2}{H_D I_T} + \frac{\sigma_{\varepsilon T}^2}{H_D m_T} + \frac{\sigma_{\varepsilon C}^2}{H_D\ m_C}\right) \quad \text{(A.8)}$$

The additional variability due to differential treatment impact will often be small in practice; in many cases the model in Equation (4.3) can be refit without the terms $\varphi_d T_{dh}$. In some settings, a differential treatment effect may occur because some districts are more supportive of interventions than others. But when the differential treatment impact arises because of unmeasured differences among students or schools, we expect that in general, the larger the topmost experimental unit in the hierarchy (in this case, districts), the smaller the value of $\sigma_\varphi^2$. This is because students tend to be more heterogeneous in larger experimental units: schools are more heterogeneous than classrooms, districts are more heterogeneous than schools, states are more heterogeneous than districts, etc. Thus, as the unit size grows, all of the variance components associated with that unit tend to decrease.

## A.5 Cross-Nested Designs

Cross-nested designs can be expressed in the general mixed model framework. In this section, we illustrate how to express the model in section 4.2.3 using the notation in this appendix. As in section A.3, suppose there are $H$ schools, with the control schools labeled as 1 to $H_C$ and the treatment schools labeled as $H_C + 1$ to $H$. For a cross-nested design, however, the ICs may be formed from students in multiple treatment schools. This is in contrast to the nested model in section A.3, where the each IC contained students from exactly one treatment school.

Consider the situation where there are a total of $I_T$ ICs. The random effects in the vector $\boldsymbol{\gamma}$ consist of the school-level effects $\xi_h$ for the treatment and control groups and the IC-level effects $\theta_i$, with respective variances $\sigma^2_{\xi T}$, $\sigma^2_{\xi C}$ and $\sigma^2_\theta$. For this cross-nested design, define

$$\boldsymbol{\gamma} = \left(\xi_1, \ldots, \xi_{H_C}, \xi_{H_C+1}, \ldots \xi_H, 0, \theta_1, \theta_2 \ldots, \theta_{I_T}\right)'.$$

This vector of random effects lists the $H$ school effects, followed by 0 to represent the IC effect for the control schools and then followed by the $I_T$ IC effects. As in section A.3, set $\boldsymbol{h_l} = (0,0,\ldots,0,1,0,\ldots,0)'$ to be the $H$-vector with a 1 in position $h$ if student $l$ is in school $h$. To account for the IC effects in the treatment (but not control) group, let $\boldsymbol{c_l} = (1,0,0,\ldots,0)'$ if student $l$ is in the control group and $\boldsymbol{t_l} = (0,0,\ldots,0,1,0,\ldots,0)'$, with a 1 in position $(i+1)$ and zeroes elsewhere, if student $l$ is in IC $i$ in the treatment group. Then $\boldsymbol{z_l} = (\boldsymbol{h_l'}, \boldsymbol{c_l'})'$ if student $l$ is in the control group and $\boldsymbol{z_l} = (\boldsymbol{h_l'}, \boldsymbol{t_l'})'$ if student $l$ is in the treatment group. Since it is assumed that all variance components are independent, for this model $\boldsymbol{G}$ is a $\left(H+1+I_T\right)\times\left(H+1+I_T\right)$ diagonal matrix with diagonal entries $\sigma^2_{\xi C}$ (repeated $H_C$ times), $\sigma^2_{\xi T}$ (repeated $H_T$ times), 0, and $\sigma^2_\theta$ (repeated $I_T$ times).

Because each IC is formed from students originating from multiple treatment schools, however, the covariance matrix $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$ is not necessarily block diagonal. The response of a student in the treatment group is correlated with the responses of other students in his or her school, and is also correlated with the responses of other students in his or her IC. The students in the IC, however, are drawn from multiple schools in the cross-nested design. As a consequence, the covariance matrix for the treatment group students can have a positive entry in any position, depending on how students are assigned to ICs.

**This page left blank for double-sided copying.**

# Appendix B

# Degrees of Freedom for PN-RCTs

**This page left blank for double-sided copying.**

# Appendix B
# Degrees of Freedom for PN-RCTs

In large studies, degrees of freedom are immaterial for the conclusions of the study. For large degrees of freedom, the *t* distribution is very close to the normal distribution. For example, the 0.05-level critical value for the normal distribution is 1.96, while the 0.05-level critical value for the *t* distribution with 40 degrees of freedom is 2.02; hypothesis tests and confidence intervals formed using the t distribution will be practically identical to those formed using the normal distribution. For small studies, however, inferences can differ depending on how the degrees of freedom are calculated. We first give general guidelines to degrees of freedom and then discuss adjustments to those degrees of freedom that arise from considering unequal variances in the treatment and control groups.

Note that most rigorous, well-powered education RCTs have sufficiently large sample sizes to justify the use of the normal distribution as an approximation to the *t* distribution for hypothesis testing. However, the number of degrees of freedom can sometimes be small (that is, less than 40) when schools or classrooms are the units of random assignment, and in that case, degrees of freedom should be considered when performing tests or constructing confidence intervals.

The general guideline for degrees of freedom in hierarchical models is to use

df = (number of independent units) – (number of parameters estimated),

where the number of parameters estimated includes the intercept, the treatment status indicator, and all baseline covariates (including block effects). In a two-sample $t$ test with all student observations independent, this guideline results in using $n-2$ df, where $n$ is the total number of students. By contrast, for a C-RCT design with a total of $n$ students nested in $H$ schools with half of the schools receiving the control protocol and the other half receiving the treatment protocol, the guideline indicates that df = $H-2$ because there are only $H$ independent units, corresponding to the schools, in the data. In a blocked design with $H$ schools, where half of the $m$ students in each school receive the treatment and the other half receive the control, we have $(m-2)$ degrees of freedom in each block, and the guideline results in $H(m-2)$ df.

The simple guideline, however, assumes that the variance for each independent unit is the same. This is not the case in PN-RCTs because the independent units in the control group often have a

different variance than the independent units in the treatment group due to the ICs. We (and SAS) use Satterthwaite's (1946) method to adjust the degrees of freedom for the unequal variances and illustrate the method for the basic PN-RCT design. The adjustment is similar for the other designs. The method also will adjust the df to account for estimating other covariates; a formula that may be used when other covariates are present is given by Fai and Cornelius (1996).

For the basic PN-RCT design, the problem of testing the ATE may be viewed as a special case of the Behrens-Fisher problem (see Dudewicz, Ma, Mai, and Su 2007 for a discussion) for conducting a hypothesis test of the difference between two treatment means with unequal variances in the two groups. Using the notation in section 3.2, if each IC in the treatment group has the same number of students, $J$, and if the control group has $n_C$ students, then the estimated treatment effect $\hat{\beta}_1$ may be written as the difference between the mean of the students in the treatment group and the mean of the students in the control group:

$$\hat{\beta}_1 = \bar{y}_T - \bar{y}_C = \frac{1}{n_T}\sum_{i=1}^{I_T}\sum_{j=1}^{J} y_{ij} - \frac{1}{n_C}\sum_{j=1}^{n_C} y_{0j} = \frac{1}{I_T}\sum_{i=1}^{I_T} \bar{y}_{Ti} - \frac{1}{n_C}\sum_{j=1}^{n_C} y_{0j}$$

The problem may be viewed as having $n_C$ independent observations in the control group, each with variance $v_C = \sigma^2_{\varepsilon C}$, and $I_T$ independent observations in the treatment group, each with variance $v_T = \sigma^2_\theta + \frac{\sigma^2_{\varepsilon T}}{J}$. With this setup, Satterthwaite's (1946) approximation to the degrees of freedom is

$$df = \frac{\left(\frac{v_C}{n_C} + \frac{v_T}{I_T}\right)^2}{\frac{v_C^2}{n_C^2(n_C - 1)} + \frac{v_T^2}{I_T^2(I_T - 1)}}$$

The df provided by this formula lies between the smaller value $(I_T - 1)$ and the sum $(n_C + I_T - 2)$.

For the example in section 3.3, we estimate $v_C$ by 197.13 and $v_T$ by (52.48 + 204.32/5) = 93.34. The Satterthwaite formula then gives 46.9 degrees of freedom, which is the value that SAS used for the test.

# Appendix C

# Analyzing PN-RCT Data Using R Software

**This page left blank for double-sided copying.**

# Appendix C
# Analyzing PN-RCT Data Using R Software

The examples in chapter 3 are all analyzed using SAS software. Using the same software package for all examples in chapter 3 allows the design and analysis features to be easily compared, without forcing the reader to switch between different output formats. Not all readers will want to use SAS for analysis, however. This appendix presents annotated commands and selected output for the designs in sections 3.3, 3.5, and 4.1 for use with the R statistical software package (R Core Team, 2013).

The contributed R package lme4 (Bates, Maechler, and Bolker 2012) may be used to fit mixed models in R. The package lme4 is constructed to allow a wide range of potential covariance structures and thus can be adapted for use with PN-RCTs. The main function in the package for fitting linear mixed models is *lmer*, and we shall use that exclusively. We assume the reader is familiar with fitting linear models in R using the function *lm*. Note that since lme4 is a contributed package to open-source software, features are subject to change in future versions. The code presented below was developed for use with version 2.15.2 of R and with version 0.999999-0 of lme4.

## C.1 Basic PN-RCT Design (Sections 3.2 and 3.3)

We first demonstrate a method for analyzing data from the basic PN-RCT design that works in R or any package that will fit a hierarchical model. We estimate the mean and variance of each group separately and then find the p-value for a large-sample test using those means and variances. The variable names used for this analysis are described in table 3. The data are assumed to be in a data frame named *model1*. The following code could alternatively be written as a function.

```
library(lme4) # load the lme4 package
model1$ic = factor(model1$ic) # Declare ic as a factor

# Find the mean and variance of the mean: control group

controlfit = lm(y ~ 1, subset=(trt==0), data=model1)
controlmean = controlfit$coef

controlmean_var = summary(controlfit)$sigma^2/
                  length(model1$y[model1$trt==0])

# Find the mean and variance of the mean: treatment group

treatfit = lmer(y ~ 1 + (1 | ic), subset=(trt==1), data=model1,
                REML=TRUE)
treatmean = attr(treatfit,"fixef")
treatmean_var = diag(vcov(treatfit))

# Combine the estimates from the treatment and control groups
# and perform the test

diff_mean = treatmean - controlmean
diff_var = treatmean_var + controlmean_var
t_stat = diff_mean/sqrt(diff_var)
p_value = (1 - pnorm(t_stat))*2 # can use t value with Satterthwaite
```

The object *treatfit* gives the variance components for the treatment group, and the objects *t_stat* and *p-value* contain the test statistic and p-value for the test of significance for the ATE. The values of these objects from R are:

```
> treatfit
Linear mixed model fit by REML
Formula: y ~ 1 + (1 | ic)
   Data: model1
 Subset: (trt == 1)
  AIC  BIC logLik deviance REMLdev
 1042 1051 -518.1     1039    1036
Random effects:
 Groups   Name        Variance Std.Dev.
 ic       (Intercept)  52.485   7.2446
 Residual             204.324  14.2942
Number of obs: 125, groups: ic, 25

Fixed effects:
            Estimate Std. Error t value
(Intercept)  105.078      1.932   54.38
> diff_mean
(Intercept)
   5.161021
> diff_var
[1] 5.310493
> t_stat
(Intercept)
   2.239588
> p_value
(Intercept)
 0.02511767
```

This p-value was calculated using a normal approximation, but the Satterthwaite degrees of freedom could be used instead with a t distribution.

If you are willing to assume that $\sigma^2_{\varepsilon C} = \sigma^2_{\varepsilon T}$, the following command in R will fit the PN-RCT in Equation (3.3). This command fits the IC effect only within the treatment group but assumes a common residual variance. Covariates can be added to this model after the '~' in the calling statement.

```
lmefit_model1 = lmer(y ~ trt + (0 + trt | ic), data=model1, REML=TRUE)
```

The object *lmefit_model1* gives the estimated variance components and treatment effect.

```
> lmefit_model1
Linear mixed model fit by REML
Formula: y ~ trt + (0 + trt | ic)
   Data: model1
  AIC  BIC logLik deviance REMLdev
 2056 2070  -1024     2054    2048
Random effects:
 Groups   Name Variance Std.Dev.
 ic       trt   53.281   7.2994
 Residual      200.343  14.1543
Number of obs: 250, groups: ic, 26

Fixed effects:
            Estimate Std. Error t value
(Intercept)   99.917      1.266   78.92
trt            5.161      2.310    2.23

Correlation of Fixed Effects:
    (Intr)
trt -0.548
```

The model fit in R gives the same parameter estimates as a SAS model that assumes the student-level variance components are the same in the control and treatment groups. The treatment effect is estimated as $\hat{\beta}_1 = 5.16$ with standard error 2.31. The variance parameter estimates, from the output, are $\hat{\sigma}_\theta^2 = 53.3$, and $\hat{\sigma}_\varepsilon^2 = 200.3$.Note that R gives a standard deviation behind each estimated variance component. This is the square root of the variance component (for example, 7.2994 is the square root of 53.281), *not* a standard error for the variance component. R does not give standard errors for the variance components. Standard errors are usually based on a normal approximation to the distribution of the statistic. Estimated variance parameters generally have a skewed distribution so the normal approximation performs poorly.

## C.2 The Blocked PN-RCT Design (Section 3.5).

For the blocked design, we cannot fit the control and treatment groups separately and then combine the variances as we did for the basic PN-RCT design in section C.1. However, R will fit a unified model in which the residual student-level variance is assumed to be the same for the treatment and control groups. The ICs must be uniquely identified across all of the schools for the model to be fit correctly, and all random factors must be declared to be factors before using the function *lmer*.

```
model2$school = factor(model2$school)
model2$ic = factor(model2$ic)
model2$schic = interaction(model2$school,model2$ic)
lmefit_model2 = lmer(y ~ trt + (1 + trt | school) + (0 + trt | schic),
                data=model2, REML=TRUE)
```

The following output shows $\hat{\beta}_0 = 102.212$ with standard error 1.457. The treatment effect is estimated as $\hat{\beta}_1 = 5.461$ with standard error 2.356. The variance parameter estimates, from the output, are $\hat{\sigma}^2_{\xi} = 15.6241$, $\hat{\sigma}^2_{\eta} = 48.1926$, $\hat{\sigma}^2_{\theta} = 5.2278$, and $\hat{\sigma}^2_{\varepsilon} = 162.0988$. The *VarCorr* command will extract the covariance matrix at each level in the hierarchy, and the $school component gives the school-level covariance parameter estimate of $\hat{\sigma}_{\xi\eta} = -9.84$.

```
> lmefit_model2
Linear mixed model fit by REML
Formula: y ~ trt + (1 + trt | school) + (0 + trt | ic2)
   Data: model2
  AIC  BIC logLik deviance REMLdev
 2416 2442  -1201     2408    2402
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 ic2      trt           5.2278  2.2864
 school   (Intercept)  15.6241  3.9527
          trt          48.1926  6.9421  -0.359
 Residual             162.0988 12.7318
Number of obs: 300, groups: ic2, 31; school, 15

Fixed effects:
            Estimate Std. Error t value
(Intercept)  102.212      1.457   70.16
trt            5.461      2.356    2.32

Correlation of Fixed Effects:
    (Intr)
trt -0.506
>
> VarCorr(lmefit_model2)
$ic2
         trt
trt 5.227778
attr(,"stddev")
     trt
2.286433
attr(,"correlation")
    trt
trt   1

$school
            (Intercept)       trt
```

```
(Intercept)   15.624131 -9.839705
trt           -9.839705 48.192641
attr(,"stddev")
(Intercept)         trt
   3.952737    6.942092
attr(,"correlation")
            (Intercept)        trt
(Intercept)   1.0000000 -0.3585864
trt          -0.3585864  1.0000000

attr(,"sc")
[1] 12.7318
```

## C.3 The Clustered Design (Sections 4.1.1 and 4.1.2)

As with the basic PN-RCT, for the clustered design we can fit the treatment and control groups separately and then combine the information if unequal variances are desired, or we can use one *lmer* command if the school-level and residual variances are assumed equal for the two groups.

Commands that may be used to fit the control and treatment schools separately, and then combine the results, are:

```
# Control group

controlfit=lmer(y~1 + (1 | school), subset=trt==0, data=model3)
controlmean = attr(controlfit,"fixef")
controlmean_var = diag(vcov(controlfit))

# Treatment group

treatfit = lmer(y ~ 1 + (1 | school/ic), subset=trt==1, data=model3,
REML=TRUE)
treatmean = attr(treatfit,"fixef")
treatmean_var = diag(vcov(treatfit))

# Combine the estimates from the treatment and control groups

diff_mean = treatmean - controlmean
diff_var = treatmean_var + controlmean_var
t_stat = diff_mean/sqrt(diff_var)
p_value = (1 - pt(t_stat,68))*2
```

The following command will fit the unified model, assuming a common variance at the student level. Note that to fit this model in R, the ICs must have unique labels for all ICs in the study. Covariates can be added to this model after the '~' in the calling statement.

```
model3$trtname = factor(model3$trt)
model3$schic = interaction(model3$school,model3$ic)
lmefit_model3 = lmer(y ~ trt + (0+ trtname | school) +
                 (0 + trt| schic), data=model3, REML=TRUE)
```

```
> lmefit_model3
Formula: y ~ trt + (0 + trtname | school) + (0 + trt | schic)
   Data: model3
   AIC   BIC logLik deviance REMLdev
 23393 23435 -11690    23384   23379
Random effects:
 Groups   Name     Variance Std.Dev. Corr
 schic    trt       31.486   5.6112
 school   trtname0  38.231   6.1831
          trtname1  58.642   7.6578  0.000
 Residual          226.810  15.0602
Number of obs: 2800, groups: schic, 175; school, 70

Fixed effects:
            Estimate Std. Error t value
(Intercept)   99.322      1.120   88.68
trt            1.422      1.821    0.78

Correlation of Fixed Effects:
    (Intr)
trt -0.615
```

The treatment effect is estimated a $\hat{\beta}_1 = 1.42$ with standard error 1.82. The variance parameter estimates, from the output, are $\hat{\sigma}^2_{\xi C} = 38.2$, $\hat{\sigma}^2_{\xi T} = 58.6$, $\hat{\sigma}^2_{\theta} = 32.3$, and $\hat{\sigma}^2_{\varepsilon} = 226.8$.

## C.4 Randomization of Schools Within Blocks (Section 4.1.3)

The design in which schools are randomly assigned to treatments within blocks is fit similarly to the clustered design in R. The following commands may be used:

```
model4$trtname = factor(model4$trt)
model4$dist = factor(model4$dist)
model4$school = factor(model4$school)
model4$schdist = interaction(model4$school,model4$dist)
model4$icschdist = interaction(model4$ic,model4$school,model4$dist)
```

```
lmefit_model4 = lmer(y ~ trt + (1 + trt | dist ) + (1 | schdist) +
                (0 + trt | icschdist), data=model4, REML=TRUE)
```

The output is interpreted in the same way as in section C.3: the command `VarCorr(lmefit_model4)` will give the covariance of the district-level effects.

# Appendix D

# Analyzing Basic PN-RCTs Using HLM Software

**This page left blank for double-sided copying.**

# Appendix D
# Analyzing Basic PN-RCTs Using HLM Software

We now demonstrate a method for analyzing data from a basic PN-RCT design that works in *HLM for Windows, Version 7.0* (Raudenbush, Bryk, & Congdon 2011). We do not provide methods for estimating parameters from the blocked and clustered designs in sections 3.5 and 4.1, and recommend using SAS or R for those designs. HLM software (as of Version 7) is not readily amenable for the blocked and clustered designs because it is designed for data in which the hierarchical variance structure is the same in the treatment and control groups. Although we examined potential ways to work around the HLM software limitations for the more complex PN-RCTs, none of these approaches calculated standard errors correctly.

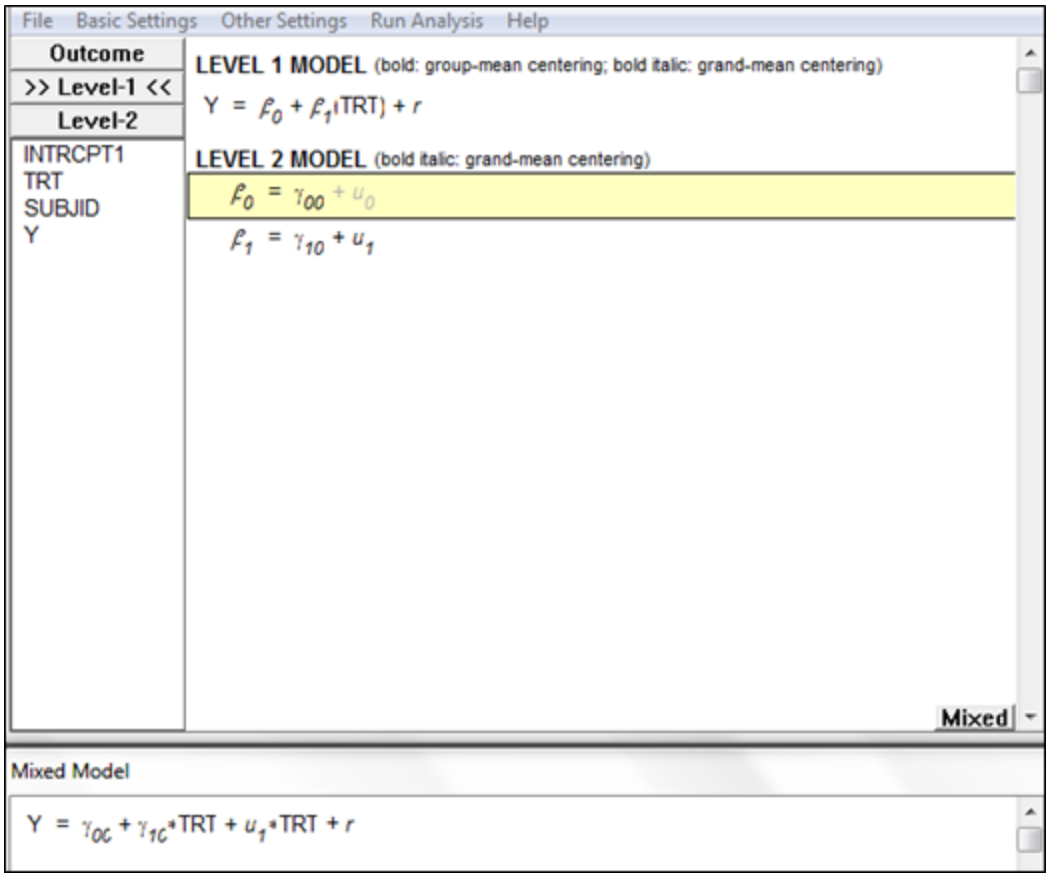
The following discussion is for the basic PN-RCT design only. The variable names used for this analysis are described in table 3. The data are assumed to be in a data frame named *model1*. There are three important notes regarding using HLM for analyzing PN-RCT data that will be discussed in turn.

First, the cluster identification (IC) values for the control condition should be renumbered as if each control is a singleton cluster, instead of using IC=0 for all controls. Using the *model1* data, for example, the first control subject would be coded as 26 in the IC column (after the 25$^{th}$ treatment IC), the second control subject would be coded 27, the third coded as 28, and so forth through 125. Coding in this manner (rather than using IC=0 for all control subjects) will not change the variance component estimates. Rather, it assists the software in treating each control as an independent unit rather than as one large cluster for df estimates. The model to be estimated in HLM is as follows.

| | |
|---|---|
| Level 1 Model: | $Y = \beta_0 + \beta_1(TRT) + r$ |
| Level 2 Model: | $\beta_0 = \gamma_{00}$ |
| | $\beta_1 = \gamma_{10} + u_1$ |
| Mixed Model: | $Y = \gamma_{00} + \gamma_{10}(TRT) + u_1(TRT) + r$ |

Because TRT is coded 1=treatment and 0=control, $\gamma_{00}$ is the expected value of the control condition, $\gamma_{10}$ is the expected value of the treatment effect, $u_1$ the treatment cluster error, and $r$ is the residual error. Although TRT is used as a "Level 1" predictor in the HLM framework, in "Level 2" the variance of the intercept will be specified as fixed while the variance of the slope is specified as random. This specification allows HLM software to limit estimation of the IC variance component

to the treatment condition. Once the data are imported into HLM2 using procedures specified in the software manual (i.e., create a new MDM file using ASCII or other software files for Level 1 and Level 2 datasets), specify Y as the outcome, add TRT as a Level 1 predictor (uncentered), turn *off* the Level 2 random intercept (clicking on $u_0$ in the $\beta_0$ line), and turn *on* the Level 2 random slope (clicking on $u_1$ in the $\beta_1$ line). The window should look as follows.




In Basic Settings, one can tailor output file names as desired; in Other Settings, one can select Estimation Settings and REML or ML, depending on preference. To save the command file, select File and Save As (browse and name the file). Finally, click Run Analysis. Below, the relevant output using REML estimation is given in the order in which the output is displayed.

```
The outcome variable is        Y
Final estimation of fixed effects:
----------------------------------------------------------------------------
                                Standard             Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f. P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          99.917357   1.265972    78.925        99    0.000
For      TRT slope, B1
   INTRCPT2, G10           5.161021   2.310760     2.233       149    0.027
----------------------------------------------------------------------------
```

The mean for the control condition is estimated as $\hat{\beta}_0 = 99.9$ with standard error 1.3. The df were computed as the total number of subjects (250) – the sum of the control subjects and treatment clusters (150) – the number of predictors (1) = 99. The treatment effect is estimated as $\hat{\beta}_1 = 5.16$ (*SE*=2.31), with a *t*-test value of 2.23 based on *df*=149, giving a p-value of 0.027. The *df* for the treatment slope was calculated as the sum of the number of control subjects and treatment clusters (150) – number of predictors (1) = 149. Note that the *df* estimates in *HLM* software correspond with *SAS* software's *ddfm*=BW option.

```
The outcome variable is        Y
Final estimation of fixed effects
(with robust standard errors)
----------------------------------------------------------------------------
                                  Standard             Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f. P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          99.917357   1.250779    79.884        99    0.000
For      TRT slope, B1
   INTRCPT2, G10           5.161021   2.269156     2.274       149    0.024
----------------------------------------------------------------------------
```

HLM next provides robust, or "sandwich," estimators of the variance; these result in the same conclusions for the tests.

```
Final estimation of variance components:
------------------------------------------------------------------------------
Random Effect           Standard      Variance     df    Chi-square  P-value
                        Deviation     Component
------------------------------------------------------------------------------
     TRT,       u1        7.30424      53.35185    24      55.91587    0.000
 level-1,       r        14.15400     200.33567
------------------------------------------------------------------------------

NOTE: The chi-square statistics reported above are based on only 25 of 150
units that had sufficient data for computation. Fixed effects and variance
components are based on all the data.

Statistics for current covariance components model
--------------------------------------------------
Deviance                        = 2048.156042
Number of estimated parameters = 2
```

The final pieces of HLM output provide the variance component estimates of $\hat{\sigma}_\theta^2 = 53.4$ and common variance for the residuals $\hat{\sigma}_\theta^2 = 200.3$, as well as the value of the deviance that can be used to evaluate the model and the number of estimated fixed effects.

For comparison, the analyses of the same data using ML estimation instead of REML (in HLM, select Other Settings, then Estimation Settings, and then click on ML) are given below. Notice that the variance components and fixed effect coefficients' standard errors are slightly smaller than with

REML estimation. Finally, notice that the deviance (-2LL, given at the end of the output) is slightly larger, and is based on four parameter estimates (fixed and random effects) rather than two (reflecting the ML prediction of both the fixed and random effects).

```
Final estimation of fixed effects:
----------------------------------------------------------------------------
                                    Standard             Approx.
   Fixed Effect         Coefficient  Error      T-ratio   d.f. P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          99.917357   1.263179    79.100        99    0.000
For      TRT slope, B1
   INTRCPT2, G10           5.161021   2.276015     2.268       149    0.025
----------------------------------------------------------------------------

The outcome variable is        Y
Final estimation of fixed effects
(with robust standard errors)
----------------------------------------------------------------------------
                                    Standard             Approx.
   Fixed Effect         Coefficient  Error      T-ratio   d.f. P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          99.917357   1.250779    79.884        99    0.000
For      TRT slope, B1
   INTRCPT2, G10           5.161021   2.269156     2.274       149    0.024
----------------------------------------------------------------------------

Final estimation of variance components:
-----------------------------------------------------------------------------
Random Effect           Standard      Variance     df    Chi-square  P-value
                        Deviation     Component
-----------------------------------------------------------------------------
     TRT,       u1        7.05160      49.72502    24      56.16343    0.000
 level-1,       r        14.12277     199.45263
-----------------------------------------------------------------------------

NOTE: The chi-square statistics reported above are based on only 25 of 150
units that had sufficient data for computation. Fixed effects and variance
components are based on all the data.

Statistics for the current model
-------------------------------------------------
Deviance                        = 2053.598202
Number of estimated parameters = 4
```

While HLM can estimate heterogeneous residual variances for the treatment and control conditions (as recommended previously in this paper), it can only do so using a maximum likelihood (ML) algorithm. The reason is that it automatically conducts a likelihood ratio test comparing the fit of the homogeneous and heterogeneous residual variance models, a test that requires use of ML estimation (there is currently no way to turn this test off in HLM software). Note that, for smaller samples, using ML instead of REML in multilevel modeling can downwardly bias variance component estimates. Given the large sample size in the example data, there will be a negligible difference between REML and ML estimates.

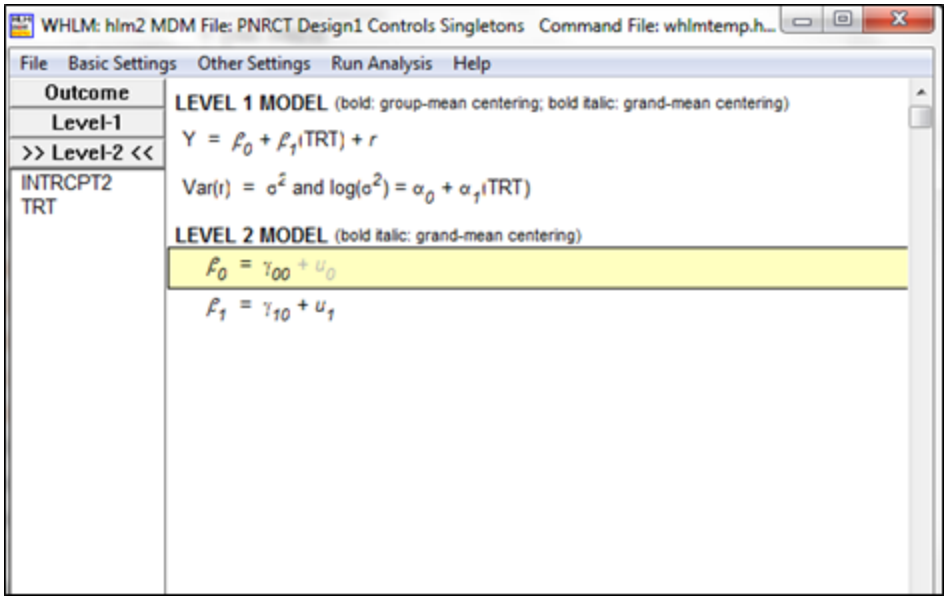
To estimate this PN-RCT model with heterogeneous residual variances for the treatment and control conditions, use the same specifications previously discussed but now select Other Settings, then Estimation Settings, then Heterogeneous Sigma^2. Next, click TRT twice and the variable will move into the box labeled "Predictors of level-1 variance." Then select "OK". The model that HLM will estimate is the same as before, except that the variance of the residual error, *r*, is restructured as follows.

Mixed Model: $Y = \gamma_{00} + \gamma_{10}(TRT) + u_1(TRT) + r$

where $Var(r) = \sigma^2$, and $\ln(\sigma^2) = \alpha_0 + \alpha_1(TRT)$

Given the restructuring of *r* and the coding of TRT, the estimated residual variance for the treatment condition is $\exp(\alpha_0 + \alpha_1*0) = \exp(\alpha_0)$, and the estimated variance for the treatment condition is $\exp(\alpha_0 + \alpha_1*1) = \exp(\alpha_0 + \alpha_1)$.

Once the model is specified in HLM, the window should look as follows.




Finally, click "Run Analysis." The results will appear in a sequence, with the homogeneous variance assumed first (same as prior output with ML estimation shown above), then the likelihood ratio test, and finally the heterogeneous variance model results. For brevity, only the likelihood ratio test and heterogeneous variance results are presented below.

```
RESULTS FOR HETEROGENEOUS SIGMA-SQUARED
(macro iteration 3)
 Var(R) = sigma^2 and
 log(sigma^2) = alpha0 + alpha1(TRT)

 Model for level-1 variance
 ----------------------------------------------------------------------
                                  Standard
    Parameter        Coefficient  Error       Z-ratio   P-value
 ----------------------------------------------------------------------
 INTRCPT1    ,alpha0     5.27585      0.126491     41.709     0.000
      TRT    ,alpha1     0.04290      0.189737      0.226     0.821
 ----------------------------------------------------------------------
```

The *HLM* output above gives the estimates of $\alpha_0$ and $\alpha_1$. Hence, the control condition residual variance is estimated to be exp(5.27585) = 195.56, and the treatment condition residual (within-cluster) variance is estimated to be exp(5.27585+0.04290) = 204.13.

```
Summary of Model Fit
 ----------------------------------------------------------------------
 Model                                Number of         Deviance
                                      Parameters
 ----------------------------------------------------------------------
 1. Homogeneous sigma_squared              4           2053.59820
 2. Heterogeneous sigma_squared            5           2053.54477
 ----------------------------------------------------------------------
 ----------------------------------------------------------------------
 Model Comparison                 Chi-square       df    P-value
 ----------------------------------------------------------------------
 Model 1 vs Model 2                   0.05343       1     >.500

tau
      TRT,B1     48.93062

Standard error of tau
      TRT,B1     26.03519

 ----------------------------------------------------
  Random level-1 coefficient   Reliability estimate
 ----------------------------------------------------
       TRT, G1                      0.545
 ----------------------------------------------------

NOTE: The reliability estimates reported above are based on only 25 of 150
units that had sufficient data for computation. Fixed effects and variance
components are based on all the data.
The value of the log-likelihood function at iteration 2 = -1.026772E+003
```

The likelihood ratio test comparing the heterogeneous and homogeneous residual variance models has p-value > 0.5, indicating that for these data, the extra complication of modeling heterogeneous variances is not necessary. The estimated IC-level variance is 48.9. The other parts of the HLM output, giving the fixed effects estimates using ML estimation, are similar to the results displayed above.

The third and final issue in using HLM software is that it does not calculate Satterthwaite (1946) *df* for heterogeneous residual variances. However, the calculations described in Appendix B can be used to adjust the *df* if needed.

```
The outcome variable is        Y
Final estimation of fixed effects:
----------------------------------------------------------------------------
                                    Standard             Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f. P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          99.917357   1.250779    79.884        99    0.000
For      TRT slope, B1
   INTRCPT2, G10           5.161021   2.270461     2.273       149    0.024
----------------------------------------------------------------------------

The outcome variable is        Y
Final estimation of fixed effects
(with robust standard errors)
----------------------------------------------------------------------------
                                    Standard             Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f. P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          99.917357   1.250779    79.884        99    0.000
For      TRT slope, B1
   INTRCPT2, G10           5.161021   2.269156     2.274       149    0.024
----------------------------------------------------------------------------

Final estimation of variance components:
------------------------------------------------------------------------------
Random Effect           Standard      Variance     df    Chi-square  P-value
                        Deviation     Component
------------------------------------------------------------------------------
     TRT,       u1        6.99504      48.93062    24      54.87687    0.001
------------------------------------------------------------------------------

NOTE: The chi-square statistics reported above are based on only 25 of 150
units that had sufficient data for computation. Fixed effects and variance
components are based on all the data.

Statistics for the current model
-------------------------------------------------
Deviance                        = 2053.544767
Number of estimated parameters = 5
```

**This page left blank for double-sided copying.**

# Appendix E

# Full SAS Code for Examples

**This page left blank for double-sided copying.**

# Appendix E
# Full SAS Code for Examples

```
/*************************************************************/
/*****************BASIC PN-RCT DESIGN*************************/
/*************************************************************/

/* Generates data and computes parameter estimates for the basic PN-RCT
   design in Section 3.2. Code was developed using SAS Version 9.3.
   Students randomly assigned to treatment or control: no clustering except
   for ICs. */

/* Initialize parameters for generating data */

%let ncontrol = 125;      /* number of students in control arm */
%let nclustrt = 25;       /* number of ICs in treatment arm */
%let icsize = 5;          /* number of students in each IC */
%let trtmean = 106;       /* mean for treatment group */
%let controlmean = 100;   /* mean for control group */
%let sig2C = 15**2;       /* sigma^2 at student level for control group */
%let sig2T = 15**2;       /* sigma^2 at student level for treatment group */
%let sig2theta = 25;      /* sigma^2_theta:  IC-level variance component */
%let numicm1 = %eval(&nclustrt - 1);
%let binary_cut = 115;    /* cutoff value for creating binary variable */

/* Generate data set */

data model1 (drop=u controlsd trtsd sigma_theta j);
   call streaminit(20850);
   retain trt u controlsd trtsd sigma_theta;
   controlsd = sqrt(&sig2C);
   trtsd = sqrt(&sig2T);
   sigma_theta = sqrt(&sig2theta);
   do ic = 1 to &nclustrt;
       trt = 1;
         u = rand('normal',&trtmean,sigma_theta);
       do j = 1 to &icsize;
           y = rand('normal',0,trtsd) + u;
             if y > &binary_cut then ybin = 1;
             else ybin = 0;
           subjid =  (ic- 1)*&icsize + j;
             subjidnest = j;
             icnest = ic;
             trtname = "Treatment";
           output;
          end;
      end;
   do subjid = &nclustrt*&icsize + 1 to &nclustrt*&icsize + &ncontrol;
       trt = 0;
         trtname = "Control";
         y = rand('normal',&controlmean,controlsd);
         if y > &binary_cut then ybin = 1;

         else ybin = 0;
```

```
         ic=0;
         icnest = subjid - &nclustrt*&icsize;
         subjidnest = icnest;
         output;
      end;

/* Plot the data */

/* Start with side-by-side boxplots for control, treatment groups */
/********************* Figure 8 *********************************/

proc sgplot data=model1;
   vbox y / category=trtname;
   yaxis label="Score";
   xaxis label = "Treatment Group" ;
run;

/* Sort the ICs by median in preparation for Figure 9 */

proc means data= model1 noprint nway;
  class trt ic;
  var y;
  output out= icmedian (drop= _type_ _freq_) median= icmedian;
run;

proc sort data=icmedian;
   by trt descending icmedian;

data icmedian;
   set icmedian;
   if trt = 0 then icval = 0;
   if trt = 1 then icval = _n_;
run;

proc sort data=model1;
  by trt ic;
proc sort data=icmedian;
  by trt ic;
data icplot;
   merge model1 icmedian;
   by trt ic;

proc sort data=icplot;
   by trt icmedian;

/* Plot ICs in ascending order (since we sorted by icmedian) */
/********************* Figure 9 *********************************/

proc sgplot data=icplot noautolegend ;
   vbox y / category=trtname group=ic grouporder=data meanattrs =
(symbol=Diamond color=black) outlierattrs = (symbol=Circle color=gray)
medianattrs = (color=black) whiskerattrs = (pattern=Solid color=gray)
lineattrs = (pattern=Solid color=gray) fillattrs = (color=lightgray);
   yaxis label="Score";
   xaxis label="Treatment Group";
run;
```

```
/* Perform simple alternative analysis appropriate for balanced data, from
Section 3.3 */

/* Calculate IC means for treatment group */

proc means data= model1 noprint nway;
  class trt ic;
  var y;
  output out= mean_ic (drop= _type_ _freq_) mean= icmean;
run;

/* Set ygrp = y for control students, ygrp = IC mean for treatment group */
data control_icmeans;
   set model1 mean_ic;
   if (icmean ne . and trt = 1) or (icmean = . and ic=0);
   ygrp = y;
   if trt = 1 then ygrp = icmean;

proc sort data=control_icmeans;
   by descending trt ;

proc ttest data=control_icmeans order=data;
   class trt;
   var ygrp;
run;

/* Perform 'first alternative analysis' in Section 3.3. This is also used to
   obtain initial values for parameters used in Figure 10. */

proc sort data=model1;
   by trt;
/* gives initial value for residual variance of control group; ignore output
for trt=1 */
/* also gives standard error for ybar_C of control group */
proc mixed data=model1;
   by trt;
   model y= /solution;
run;

/* gives initial value for IC and residual variance of treatment group;
ignore output for trt=0 */
/* also gives estimated standard error for ybar_T of treatment group */
proc mixed data=model1;
   by trt;
   model y=/solution;
   random intercept/subject=ic;
run;

/************************Code in Figure 10 ******************************/
/* Using ic = 0 for control group */

proc mixed data=model1 ;
  class subjid trtname ic;
  model y=trt / solution ddfm=sat ;
  random intercept / group=trtname subject=ic(trtname);
  parms (0) (52) (197) (204) / hold = 1;
```

```
  repeated subjid/ group=trtname;
  title 'Random Effects Estimator';
run;

/* The following commands give variations of the model */

/* Fit model, including regression diagnostics */

proc mixed data=model1;
  class subjid trt ic;
  model y=trt / solution ddfm=sat influence residual;
  random intercept / group=trt subject=ic(trt);
  parms (0) (52) (197) (204) / hold = 1;
  repeated subjid/ group=trt;
run;

/* Fit the models using maximum likelihood, to test whether the residual
variances are equal */

proc mixed data=model1 method=ml; /* with both variances */
  class subjid trtname ic;
  model y=trt / solution ddfm=sat ;
  random intercept / group=trtname subject=ic(trtname);
  parms (0) (52) (197) (204) / hold = 1;
  repeated subjid/ group=trtname;
run;

proc mixed data=model1 method=ml; /* with common residual variances */
  class subjid trtname ic;
  model y=trt / solution ddfm=sat ;
  random intercept / group=trtname subject=ic(trtname);
  parms (0) (52) (197)  / hold = 1;
run;

/*************************************************************/
/***************BLOCKED PN-RCT DESIGN*************************/
/*************************************************************/

/* Generates data and computes parameter estimates for Section 3.5
   Students randomly assigned to treatment or control within each school;
   treatment students randomly assigned to ICs. Schools are blocking factor
   in this design. */

/* Initialize parameters for generating data */

%let nschool = 15; /* number of schools */
%let ncontrol = 10;  /* number of control students in each school */
%let nclustrt = 2;  /* number of ICs in each school */
%let icsize = 5;     /* number of students in each IC */
%let trtmean = 106;    /* mean for treatment group */
%let controlmean = 100;  /* mean for control group */
%let sig2C = 175;   /* sigma^2_student in control group */
%let sig2T = 175;  /* sigma^2_student in treatment group */
%let sig2theta = 25;  /* variance component for ICs */
%let sig2xi = 45; /* variance component for schools  */
%let sig2eta = 10; /* variance component for random school slope */
%let sigxieta = 0;  /* covariance of school intercept and slope */
```

```
data model2 (drop=u controlsd trtsd schoolint schoolslope sigma_theta
sigma_eta sigma_xi rhoschool j);
   call streaminit(02134);
   retain trt trtname u controlsd trtsd sigma_theta sigma_eta sigma_xi ;
   controlsd = sqrt(&sig2C);
   trtsd = sqrt(&sig2T);
   sigma_theta = sqrt(&sig2theta);
   sigma_xi = sqrt(&sig2xi);
   sigma_eta = sqrt(&sig2eta);
   if sigma_eta = 0 or sigma_xi = 0 then rhoschool = 0;
   else rhoschool = &sigxieta/( sigma_xi*sigma_eta);
   do school = 1 to &nschool;
     schoolint = rand('normal',0,1);
/*generate correlated variables, then multiply by sd's*/
       schoolslope = rhoschool*schoolint +
            sqrt(1-rhoschool**2)*rand('normal',0,1);
       schoolint = schoolint*sigma_xi;
       schoolslope = schoolslope*sigma_eta;

       /* control students */

     do j = 1 to &ncontrol;
         trt = 0;
         trtname = 'Control   ';
             ic = 0;
         y = rand('normal',&controlmean,controlsd) + schoolint;
             subjnest = j;
             subjid = subjnest + (school-1)*(&ncontrol + &nclustrt*&icsize);
         output;
     end;
       /* treatment students */
       do ic = 1 to &nclustrt;
        if sigma_theta > 0 then u = rand('normal',0,sigma_theta);
            else u = 0;
        do j = 1 to &icsize;
             trt = 1;
             trtname = 'Treatment';
             y = rand('normal',&trtmean,trtsd) + u + schoolint + schoolslope;
             subjnest = j;
             subjid = subjnest + &ncontrol + (ic-1)*&icsize +
                  (school-1)*(&ncontrol + &nclustrt*&icsize);
             output;
        end;
      end;
    end;
run;

data model2put;
   set model2;
   file model2 delimiter=',';
   put trt school ic subjid y;
run;

/* graph the data */

/* Do boxplots of treatment/control scores in each school */
```

```
/******************** Figure 12 ******************************/
/*The code below produces the default SAS colors of blue and red.
  To obtain the grayscale plot in Figure 12, use
  %modstyle(parent=statistical,name=fig9st,type=CLM,
          LineStyles=solid,markers=circle,
          colors=gray82 gray46,fillcolors=grayBE gray6E);
  ods listing style=fig9st;   */

proc sgplot data=model2;
   vbox y / category=school group=trtname meanattrs=(symbol=Diamond)
medianattrs = (color=black);
   yaxis label= 'Score';
   xaxis label = 'School';
   title 'Boxplots of Test Scores, by School';
run;

/******************** Figure 13 ******************************/

/* Perform a t test using the 15 values of the ATEs for
   the individual schools. */

proc sort data=model2;
   by school trt;
proc means data=model2 noprint;
   by school trt;
   var y;
   output out = schmean mean = groupmean;

data schoolmean;
   set schmean;
   retain control;
   if trt = 0 then control = groupmean;
   else if trt = 1 then do;
      trt = groupmean;
        output;
        end;

proc ttest data=schoolmean;
   paired trt*control;

/******************** Figure 14 ******************************/

/* Obtain initial parameter estimates by fitting model without IC effects */

proc mixed data=model2 noclprint;
   class trtname school ic subjid;
   model y = trt/ ddfm = sat solution;
   random intercept trt/ subject=school type=un;  /* Fit random coefficient
regression model */
   title 'Blocked PN-RCT, get initial estimates for parameters';

/* Fit model with code in Figure 14 */

proc mixed data=model2 noclprint;
   class trtname school ic subjid;
   /* Model: 'solution' gives ATE; 'cl' gives confidence interval */
   model y = trt/ ddfm = sat solution cl;
```

```
   /* First random statement: Fit random coefficient regression model */
   random intercept trt/ subject=school type=un;
   /* Second random statement: Random effect of ICs,
      only for treatment students */
   random intercept/ group=trtname subject=ic(trtname school);
   /* Allow separate student-level variances */
   repeated subjid /group=trtname ;
   parms (15) (-10) (50) (0) (8) (162) (155)/ hold = 4;
   title 'Random Block PN-RCT';
run;
/* Fit model with common student-level variance for comparison with R */

proc mixed data=model2 noclprint noitprint;
   class trtname school ic subjid;
   model y = trt/ ddfm = sat solution;
   random intercept trt/ subject=school type=un;
   random intercept/ group=trtname subject=ic(trtname school);
   parms (15) (-10) (50) (0) (8) (160) / hold = 4;
run;



/*************************************************************/
/*********************CLUSTERED DESIGN ***********************/
/*************************************************************/

/* Generates data and computes parameter estimates for Section 4.1.
   Schools randomly assigned to treatment or control: students in treatment
   schools randomly assigned to different ICs*/

/* Initialize parameters for generating data */

%let nschcont = 35;  /* number of control schools */
%let nschtrt = 35;   /* number of treatment schools */
%let ncontrol = 40;  /* number of students in each control school */
%let nclustrt = 4;   /* number of ICs in each treatment school */
%let icsize = 10;     /* number of students in each IC */
%let trtmean = 103;    /* mean for treatment group */
%let controlmean = 100;  /* mean for control group */
%let sig2C = 225;   /* sigma^2_student in control group */
%let sig2T = 225;   /* sigma^2_student in treatment group */
%let sig2theta = 30;  /* variance component for ICs */
%let sig2xiC = 45;  /* variance component for schools in control group */
%let sig2xiT = 45; /* variance component for schools in treatment group */
%let binary_cut = 115;  /* cutoff value for creating binary variable */

data model3 (drop=u schooleff controlsd trtsd sigma_theta sigma_xiT sigma_xiC
j);
   call streaminit(90210);
   retain trt u controlsd trtsd sigma_theta sigma_xiT sigma_xiC schooleff trt
trtname;
   controlsd = sqrt(&sig2C);
   trtsd = sqrt(&sig2T);
   sigma_theta = sqrt(&sig2theta);
   sigma_xiT = sqrt(&sig2xiT);
   sigma_xiC = sqrt(&sig2xiC);
   do school =  1 to  &nschcont;
```

```
      schoolnest = school;
      schooleff = rand('normal',0,sigma_xiC);
      trt = 0;
        trtname = "Control   ";
      do icnest = 1 to &ncontrol;
          y = rand('normal',&controlmean,controlsd) + schooleff;
          subjid = icnest + (school-1)*&ncontrol;
          ic = 0;
          if y ge &binary_cut then ybin = 1;
          else ybin = 0;
          output;
        end;
       end;
   do school = &nschcont+1 to &nschtrt+ &nschcont;
      schoolnest = school - &nschcont;
      schooleff = rand('normal',0,sigma_xiT);
      trt = 1;
        trtname = "Treatment";
      do ic = 1 to &nclustrt;
          u = rand('normal',&trtmean,sigma_theta);
          icnest = ic;

        do j = 1 to &icsize;
            y = rand('normal',0,trtsd) + u + schooleff;
            if y ge &binary_cut then ybin = 1;
            else ybin = 0;
            subjid =  (school-1)*&nclustrt*&icsize + (ic-1)*&icsize + j;
            output;
           end;
         end;
     end;
run;

/* Plot the data */
/******************** Figure 15 ********************************/
proc sgplot data=model3;
   vbox y / category=trtname;
   yaxis label="Score";
   xaxis label = "Treatment Group" ;
   title 'Boxplots of Test Scores';
run;

/******************** Figure 16 ********************************/
/* Calculate the median value for each school */

proc means data= model3 noprint nway;
  class trt school;
  var y;
  output out= school_median (drop= _type_ _freq_) median= schmedian;

proc sort data=model3;
  by trt school;
proc sort data=school_median;
  by trt school;
data school_plot;
   merge model3 school_median;
   by trt school;
```

```
proc sort data=school_plot;
   by trt schmedian;

/* Plot ICs in ascending order (since we sorted by schmedian) */

proc sgplot data=school_plot noautolegend;
   vbox y / category=trtname group=school grouporder=data meanattrs =
(symbol=Diamond color=black) outlierattrs = (symbol=Circle color=gray)
medianattrs = (color=black thickness=2) whiskerattrs = (pattern=Solid
color=gray) lineattrs = (pattern=Solid color=gray) fillattrs =
(color=lightgray);
   yaxis label='Score';
   xaxis label='Treatment Group';
   title 'Boxplots of Test Scores for Each School';

/* Check means, variances of treatment and control schools; do t test using
school means */

proc sort data=model3;
  by trt school ic subjid;

proc means data=model3 noprint;
   by trt school;
   var y;
   output out=sch_means mean = sch_mean n = sch_n;

proc univariate data=sch_means; /* Find variance among control, trt group
means */
   by trt;
   var sch_mean;

proc sort data=sch_means;
   by descending trt;

proc ttest data=sch_means order=data plots=none;
   class trt;
   var sch_mean;
   weight sch_n;

/* find initial values for proc mixed */

proc mixed data=model3;  /* Gives variance components for control group;
ignore output for trt=1 */
   by trt;
   class school;
   model y = /solution;
   random intercept / subject=school;

proc mixed data=model3;  /* Gives variance components for treatment group;
ignore output for trt=0 */
   by trt;
   class school ic;
   model y = /solution;
   random intercept / subject=school;
   random intercept / subject=ic(school);
```

```
/* assumes common school variance component */
proc mixed data=model3;
   class trtname school ic subjid;
   model y = trt / ddfm = sat solution;
   random intercept / subject=school(trtname) ;
   random intercept / group=trtname subject=ic(school trtname) ;
   repeated subjid / group=trtname ;
   parms (40) (0) (30) (230) (250)/ hold = 2;

/* allows school variance components to differ */
/************************* Figure 17 ******************************/
proc mixed data=model3 noclprint;
   class trtname school ic subjid;
   model y = trt / ddfm = sat solution;
   random intercept / group=trtname subject=school(trtname) ;
   random intercept / group=trtname subject=ic(school trtname) ;
   repeated subjid / group=trtname ;
   parms (38) (58) (0) (31) (221) (233)/ hold = 3;
run;


/*************************************************************/
/**********Blocked design with school randomization ***********/
/*************************************************************/

/* Generates data and computes parameter estimates for the PN-RCT of Section
   4.1.3. (Output not shown in paper.) Schools randomly assigned to treatment
   or control within each district; treatment students randomly assigned to
   ICs. Districts are the blocking factor in this design,
   and schools are the randomization units. */

/* Initialize parameters for generating data */

%let n_district = 40;    /* number of districts */
%let n_control_school = 2; /* number of control schools in each district*/
%let n_treat_school = 2; /* number of treatment schools in each district*/
%let ncontrol = 40;  /* number of students in each control school */
%let nclustrt = 4;  /* number of ICs in each treatment school */
%let icsize = 10;     /* number of students in each IC */
%let trtmean = 103;    /* mean for treatment group */
%let controlmean = 100;  /* mean for control group */
%let sig2C = 225;   /* sigma^2_student in control group */
%let sig2T = 225;  /* sigma^2_student in treatment group */
%let sig2theta = 20;  /* variance component for ICs */
%let sig2xi = 30; /* variance component for schools within districts  */
%let sig2delta = 40; /* variance component for random district intercept  */
%let sig2phi = 30; /* variance component for random district slope */
%let sigdeltaphi = 0;  /* covariance of district intercept and slope */

data model4 (drop=u controlsd trtsd distint distslope sigma_theta sigma_delta
sigma_xi sigma_phi rhodist j);
   call streaminit(02134);
   retain trt trtname u controlsd trtsd sigma_theta sigma_xi ;
   controlsd = sqrt(&sig2C);
   trtsd = sqrt(&sig2T);
   sigma_theta = sqrt(&sig2theta);
   sigma_xi = sqrt(&sig2xi);
```

```
   sigma_delta = sqrt(&sig2delta);
   sigma_phi = sqrt(&sig2phi);
   if sigma_delta = 0 or sigma_phi = 0 then rhodist = 0;
   else rhodist = &sigdeltaphi/( sigma_delta*sigma_phi);
   do dist = 1 to &n_district;
     distint = rand('normal',0,1); /*generate correlated variables, then
multiply by sd's*/
       distslope = rhodist*distint + sqrt(1-rhodist**2)*rand('normal',0,1);
       distint = distint*sigma_delta;
       distslope = distslope*sigma_phi;
       /* control schools */
     do school = 1 to &n_control_school;
       schoolint = rand('normal',0,sigma_xi) + distint;

     do j = 1 to &ncontrol;
         trt = 0;
         trtname = "Control   ";
             ic = 0;
         y = rand('normal',&controlmean,controlsd) + schoolint;
             subjnest = j;
             subjid = subjnest + (dist-
1)*(&n_control_school*&ncontrol+&n_treat_school*&nclustrt*&icsize) +(school-
1)*(&ncontrol );
          output;
     end;
       end;
       /* treatment schools */
     do school = &n_control_school + 1 to &n_treat_school +
&n_control_school;
       schoolint = rand('normal',0,sigma_xi) + distint + distslope;
       do ic = 1 to &nclustrt;
        if sigma_theta > 0 then u = rand('normal',0,sigma_theta);
            else u = 0;
        do j = 1 to &icsize;
             trt = 1;
               trtname = "Treatment";
             y = rand('normal',&trtmean,trtsd) + u + schoolint ;
                 subjnest = j;
                 subjid = subjnest + (dist-1)*
(&n_control_school*&ncontrol+&n_treat_school*&nclustrt*&icsize) +
&n_control_school*&ncontrol + (ic-1)*&icsize + (school-&n_control_school-
1)*(&nclustrt*&icsize);
             output;
        end;
      end;
    end;
    end;
run;

/* Fit Mixed Model. SAS ensures the G matrix
   is positive definite, but you need to verify the positive definiteness for
   the unrestricted components of the district-level variance. */

proc mixed data=model4;
   class dist trtname school ic ;
   model y = trt/ ddfm = sat solution;
```

```
   random intercept trt/ subject=dist type=un;  /* Fit random coefficient
regression model for district */
   random intercept / subject=school(dist trtname);  /* Random effect for
school */
   random intercept/ group=trtname subject=ic(dist school trtname); /* Random
Effect of ICs- Only for Ts */
   parms (40) (-20) (12) (10) (0) (1) (238) / hold = 5;
   title 'Random Blocks (Districts) with Schools as Unit of Randomization';

proc mixed data=model4;
   class dist trt trtname school ic subjid;
   model y = dist*trt/ ddfm = sat solution;
   random intercept/ subject=school(trt dist);
   random intercept/ group=trt subject=ic(trt dist school);
   parms (40) (0) (1) (238) / hold = 2;
   title 'Fixed Block Effects (Districts) with Schools as Unit of
Randomization: Model 4';

run;
```


www.ed.gov

ies.ed.gov