2013

PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies

Rebecca A. Maynard
University of Pennsylvania, rmaynard@gse.upenn.edu

Nianbo Dong

Recommended Citation
Maynard, R. A., & Dong, N. (2013). *PowerUp!*: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies. *Journal of Research on Educational Effectiveness, 6* (1), 24-67. http://dx.doi.org/10.1080/19345747.2012.673143

PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies

Abstract
This paper complements existing power analysis tools by offering tools to compute minimum detectable effect sizes (MDES) for existing studies and to estimate minimum required sample sizes (MRSS) for studies under design. The tools that accompany this paper support estimates of MDES or MSSR for 21 different study designs that include 14 random assignment designs (6 designs in which individuals are randomly assigned to treatment or control condition and 8 in which clusters of individuals are randomly assigned to condition, with models differing depending on whether the sample was blocked prior to random assignment and by whether the analytic models assume constant, fixed, or random effects across blocks or assignment clusters); and 7 quasi-experimental designs (an interrupted time series design and 6 regression discontinuity designs that vary depending on whether the sample was blocked prior to randomization, whether individuals or clusters of individuals are assigned to treatment or control condition, and whether the analytic models assume fixed or random effects).

Keywords
sample design, power analysis, minimum detectable effect size (MDES), minimum required sample size (MSSR), multilevel experimental and quasi-experimental designs

Disciplines
Education | Educational Assessment, Evaluation, and Research

Running head: *PowerUp!* - A TOOL FOR MDES AND MRES

***PowerUp!*: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-experimental Design Studies**

Nianbo Dong and Rebecca Maynard

Contact information:

Nianbo Dong
Peabody Research Institute
Vanderbilt University
Peabody #181
230 Appleton Place
Nashville, TN 37203
Tel: (615) 343-2370
Email: dong.nianbo@gmail.com

Rebecca Maynard (Corresponding Author)
Graduate School of Education
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
Tel: (215) 898-3558
Email: rmaynard@gse.upenn.edu

Citation for Published Version: Dong, N. & Maynard, R. A. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies, *Journal of Research on Educational Effectiveness, 6*(1), 24-67. doi: 10.1080/19345747.2012.673143

Abstract

This paper complements existing power analysis tools by offering tools to compute minimum detectable effect sizes (MDES) for existing studies and to estimate minimum required sample sizes (MRSS) for studies under design. The tools that accompany this paper support estimates of MDES or MSSR for 21 different study designs that include 14 random assignment designs (6 designs in which individuals are randomly assigned to treatment or control condition and 8 in which clusters of individuals are randomly assigned to condition, with models differing depending on whether the sample was blocked prior to random assignment and by whether the analytic models assume constant, fixed, or random effects across blocks or assignment clusters); and 7 quasi-experimental designs (an interrupted time series design and 6 regression discontinuity designs that vary depending on whether the sample was blocked prior to randomization, whether individuals or clusters of individuals are assigned to treatment or control condition, and whether the analytic models assume fixed or random effects).

Key words: sample design; power analysis, minimum detectable effect size (MDES), minimum required sample size (MSSR); multilevel experimental and quasi-experimental designs

***PowerUp!*: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-experimental Design Studies**

Experimental and quasi-experimental designs are widely applied to evaluate the effects of policy and programs. It is important that such studies be designed to have adequate statistical power to detect meaningful size impacts, if they occur. Some excellent tools have been developed to estimate the statistical power of studies with particular characteristics to detect true impacts of a particular size or larger—referred to as Minimum Detectable Effect Sizes (MDES)—for both individual and group-randomized experiments (e.g., Optimal Design Version 2.0 (Spybrook, Raudenbush, Congdon, & Martinez, 2009 and Hedges & Rhoads, 2010; Konstantopoulos, 2009). This paper and the associated computational tools in the accompanying workbook, *PowerUp!*, use the framework of MDES formulae in these other tools to define and apply formulae to compute minimum detectable effect sizes under a variety of experimental and quasi-experimental study designs and to estimate the minimum required sample size to achieve a desired level of statistical power under various study designs and assumptions.

The paper begins with a discussion of the various study designs included in the *PowerUp!* tool. The second section discusses study designs and the design qualities that are associated with statistical power, minimum detectable effect sizes, and minimum required sample sizes for various study goals and designs. The third section of the paper presents a framework for selecting the minimum relevant effect size (MRES) to focus on when designing a study and defines the basic computational formulas for determining minimum detectable effect sizes, given study design parameters. The fourth section describes the use of the *PowerUp!* tools for estimating MRES; and the fifth section discusses the use of *PowerUp!* tools for estimating

minimum required sample sizes (MRSS) for studies with particular goals and design parameters.

STUDY DESIGNS

PowerUp! focuses on two broad classes of experimental designs, individual and cluster random assignment designs, and two classes of quasi-experimental designs – regression discontinuity designs and interrupted time series designs. In total, the *PowerUp!* tool covers 21 design variants, the key features of which are summarized in Table 1.

----- Insert Table 1 about here -----

Experimental Designs

Experimental design studies involve random assignment of study units to conditions, generally treatment or control. If experimental design studies are well implemented and the data are properly analyzed, they generate unbiased estimates of both the average effects of the program, policy, or practice being tested and the confidence intervals around the estimated impacts (Boruch, 1997; Orr, 1998; Murnane & Willett, 2010).

Individual random assignment (IRA) designs are the most common and simplest experimental design and involve the random assignment of individual analysis units to treatment or control conditions (see Table 1, model 1.0). These are also referred to in the literature as "completely randomized controlled trials" or "simple random assignment" designs. In cases where the treatment and control groups are equal in size, formulas found in sample design textbooks can be used for computing statistical power and minimum sample sizes needed to achieve certain minimum detectable effect sizes (e.g., see Orr 1998). However, when groups are unequal in size and when randomization has occurred among individuals within blocks or strata (i.e., blocked individual random assignment or BIRA designs), it is more complicated to find,

interpret, and apply the formulas for such computations (see Table 1, models 2.1 through 2.5).

Cluster random assignment designs have been gaining popularity in education research (Kirk 1995). These designs entail random assignment of clusters of analysis units (e.g., classes of students or whole schools of teachers) to the treatment or control condition. In the simplest case, all clusters in a study sample are randomized either individually or within "blocks" (e.g., defined by district or state), resulting in what is referred to as cluster (or group) random assignment designs. These models generally fall into one of two categories—simple cluster random assignment (CRA) designs (see Table 1, models 3.1 through 3.3) or blocked cluster random assignment (BCRA) designs (see Table 1, models 4.1-4.5). In simple cluster random assignment designs, top-level clusters (e.g., schools containing teachers and students) are randomly assigned to the treatment or control condition (e.g., see Borman, Slavin, Cheung, Chamberlain, and Madden, et al., 2007; Cook, Hunt, & Murphy, 2000). In contrast, in blocked cluster random assignment designs, sub-clusters of individuals within top-level clusters (blocks) are randomly assigned to the treatment or control condition (e.g., see Nye, Hedges, & Konstantopoulos, 1999).

In order to determine the minimum detectable effects size (MDES) for a particular sample size and allocation or the minimum required sample size (MRSS) to achieve a target MDES, it is necessary to account for both the particular qualities of the study design and the implication of that design for the analytic models to be used. For example, individual random assignment design studies typically use simple multiple regression models, whereas blocked individual random assignment designs and cluster random assignment design studies generally use hierarchical linear models (HLM) that account for clustering of the analysis units (e.g., students within classrooms or students within classrooms, schools and districts). Blocked random

assignment designs, whether individual or cluster level random assignment, typically entail meta-analyzing the results of mini-studies of each sample using a fixed or random block effect model.

PowerUp! supports computation of both MDES and MRSS for a variety of individual and cluster random assignment designs that are distinguished by whether the study sample is blocked prior to assigning units to treatment or control condition, by the number of levels of clustering, and by the level at which random assignment occurs. For example, Model 1.0 (IRA and N_IRA) entails neither blocking nor clustering, while Model 2.1 (BIRA2_1c and N_BIRA2_1c) refers to a blocked individual random assignment design that assumes constant effects across the assignment blocks. Model 3.1 (CRA2_2r and N_CRA2_2r) pertains to a design in which there are two levels of sample clustering, assignment to treatment or control condition is at the second level (e.g., students are the units for analysis, classes of students randomized to condition), and impacts are estimated using a random effects model. Model 3.3 (CRA4_4r and N_CRA4_4r) is similar to Model 3.1, except that it pertains to a design with four levels of sample clustering and random assignment occurring at the fourth level.

The suffix of the Worksheet names in *PowerU!* shown in Table 1, columns 7 and 8 denote key characteristics of the study design and intended analytic model. For example, for models 2.1 through 2.4, denoted by BIRAi_jk, *i* takes on the values of 2 through 4 to denote the levels of blocking; *j* takes on the values of 1 through 3 to denote the level at which random assignment occurs (e.g., students = 1, schools = 2, and districts or states = 3); and *k* takes on values of c, f, and r, denoting the assumptions to be used in estimating the effects of the treatment. A "c" denotes the assumption of *constant* treatment effects across blocks, an "f" denotes the assumption that the block effect is *fixed* (i.e., each block has specific treatment effect which

could differ across block), and an "r" denotes the assumption that the block effect is *random* (i.e., the treatment effects can randomly vary across blocks).

In general, the decision about whether to use a fixed block effect model or a random block effect model depends on the sampling scheme used in the study and the population to which the results will be generalized. If the study uses a random sample drawn from a population to which the results are expected to generalize, the random block effect model would be appropriate. However, if the intent is to generalize the findings only to the study sample, a fixed block effect would be appropriate, with the block indicators functioning as covariates controlling for the treatment effects of block membership. With this model, estimates of the average treatment effect and its standard error are computed by averaging the block-specific treatment effects and computing the standard error of that average, while with the random block effect model estimates one average treatment effect across all blocks and one standard error. Key properties of these models are illustrated in Table 2.

----- Insert Table 2 about here -----

The first three blocked random assignment design models in the tool kit pertain to 2-level designs. Model 2.1 (used in *PowerUp!* worksheets BIRA2_1c and N_ BIRA2_1c) assumes treatment effects are constant across blocks and that the results pertain to the population groups similar to the student sample; Models 2.2 (used in BIRA2_1f and N_BIRA2_1f) assumes that the treatment effects within blocks (for example, schools) are fixed, but they may differ across blocks, and that the estimated impacts pertain to population groups similar to the schools represented in the sample; and Model 2.3 (used in BIRA2_1r and N_BIRA2_1r) assumes that the treatment effects may vary randomly across blocks and that the estimated average effect is generalizable to the reference population for the study (for example, all students and schools).

Models 2.4 and 2.5 (used in BIRA3_1r and N_BIRA3_1r, and BIRA4_1r and N_BIRA4_1r, respectively) assume that random assignment occurs at level 1 (e.g., students) and that impacts of the treatment vary randomly across higher levels (e.g., classrooms, schools, districts).

Models 4.1 through 4.5 are counterpart blocked cluster random assignment models (denoted as BCRAi_jk and N_BCRAi_jk). Models 4.1 and 4.4 (BCRA3_2f and BCRA4_3f) assume that random assignment occurs at level 2 and level 3, respectively (e.g., school and district, respectively) and that the treatment effects are fixed across blocks, as in the case of Model 2.2 above. Models 4.2 and 4.3 are similar to models 2.4 and 2.5 above, except that the random assignment occurred at level 2. Model 4.5 is similar to model 2.5 above, except that it assumes that random assignment occurred at level 3, not level 1.

Quasi-experimental Designs

In quasi-experimental designs, comparison groups are identified by means other than random assignment (e.g., students scoring just above the cut-point on the test used to select the treatment group, which consists of those with scores below the cut-point or students in matched schools not offering the treatment). Although there is a rich literature demonstrating the limitations of quasi-experimental methods for estimating treatment effects, quasi-experimental methods will continue to be used when it is not practical or feasible to conduct a study using random assignment to form the treatment and comparison groups. Thus, *PowerUp!* includes tools for estimating MDES for studies that use two quasi-experimental—regression discontinuity designs and interrupted time series designs.

Regression discontinuity (RD) **designs** compare outcomes for the treatment group (e.g., students with low pretest scores or schools designated in need of improvement based on the percent of students scoring below proficient on a state test) with a comparison group that was

near the threshold for selection for the treatment on the basis of some characteristic that is measured using an ordinal scale (e.g. the pretest score or the percent of students scoring below proficient on the state test), but that was not selected. Under certain conditions, studies that compare groups on either side of this selection threshold will yield unbiased estimates of the local average treatment effect for individuals whose "score" on the selection criterion is in the vicinity of the selection threshold or "discontinuity" (Bloom, 2009; Cook & Wong, 2007; Imbens & Lemieux, 2008; Schochet, 2008b; Schochet, Cook, Deke, Imbens, Lockwood, Porter, & Smith, 2010; Shadish, Cook, & Campbell, 2002; Thistlethwaite & Campbell, 1960; Trochim, 1984). In recent years, RD designs have been applied to study the effects on academic achievement of a variety of policies and practices, including class size reductions (Angrist & Lavy, 1999), mandatory summer school (Jacob & Lefgren, 2004; Matsudaira, 2008), and the federal Reading First Program (Gamse et al., 2008).

For sample design purposes, RD designs can be mapped to corresponding random assignment study designs in terms of the unit of assignment to treatment and the sampling framework (Schochet, 2008b). *PowerUp!* includes tools for estimating MDES for six specific RD designs described in Table 1, above:

- Model 5.1: "Students are the unit of assignment and site (e.g., school or district) effects are fixed" (Schochet, 2008b, p.5). This corresponds to the 2-level blocked individual random assignment designs with fixed effects and treatment at level 1 (Table 1, Model 2.2).
- Model 5.2: "Students are the units of assignment and site effects are random" (Schochet, 2008b, p.5). This corresponds to 2-level blocked individual random assignment designs with random block effects (Table 1, Model 2.3).

- Model 5.3: "Schools are the unit of assignment and no random classroom effects" (Schochet, 2008b, p.5). This corresponds to 2-level simple cluster random assignment designs (Table 1, Model 3.1).
- Model 5.4: "Schools are the units of assignment and classroom effects are random" (Schochet, 2008b, p.6). This corresponds to 3-level simple cluster random assignment designs with treatment at level 3 (Table 1, Model 3.2).
- Model 5.5: "Classrooms are the units of assignment and school effects are fixed" (Schochet, 2008b, p.5). This corresponds to 3-level blocked cluster random assignment designs with fixed effects and treatment at level 2 (Table 1, Model 4.1).
- Model 5.6: "Classrooms are the units of assignment and school effects are random" (Schochet, 2008b, p.6). This corresponds to 3-level blocked cluster random assignment designs with treatment at level 2 and random effects across clusters (Table 1, Model 4.2).

Interrupted time-series (ITS) designs are used to estimate treatment impact by comparing trends in the outcome of interest prior to the introduction of the treatment and after (Bloom, 1999). They have been used primarily in large-scale program evaluations where program or policy decisions did not include or allow selecting participants or sites using a lottery. Examples include evaluations of the Accelerated Schools reform model (Bloom et al., 2001), First Things First school reform initiative (Quint, Bloom, Black, Stephens, & Akey; 2005), Talent Development (Kemple, Herlihy, & Smith; 2005), Project GRAD (Quint, Bloom, Black, & Stephens; 2005), and the Formative Assessments of Student Thinking in Reading (FAST-R) Program (Quint, Sepanik, & Smith, 2008).

A challenge with ITS designs is establishing a credible basis for determining the extent to which changes occurring after the onset of the intervention can be attributed reasonably to the

intervention rather than to other factors. One strategy for improving the ability to parse out effects of co-occurring factors that can affect observed differences in outcomes between the pre- and post-intervention period is to use both before-and-after comparisons within the time-series (e.g., schools before and after the introduction of the treatment) and comparison of the time series for the treatment units with a matched group of units that never received the treatment.

PowerUp! includes tools for estimating the MDES and the minimum sample size requirements for ITS design studies that involve up to two levels of clustering (see Table 1, Model 6.0). For example, as in the applications cited above, the treatment is often delivered at the cohort level, while the analysis is conducted at the student level, and the school is used as constant or fixed effect.

FACTORS THAT AFFECT MINIMUM DETECTABLE EFFECT SIZES AND MINIMUM REQUIRED SAMPLE SIZES

Smartly designed evaluations have sample sizes large enough that, should the program, policy, or practice under study have a meaningful size impact, there is a high probability that the study will detect it. However, knowing how large a sample is sufficient for this purpose depends on a number of factors, some of which can only be "guesstimated" prior to conducting the study. Moreover, some of these factors are discretionary (i.e., based on the evaluator's judgment) and others are inherent (i.e., depend on the nature of the intervention and the study design). Put another way, discretionary factors are statistical qualifications decided on by the evaluator, while inherent factors are characteristics of the true effect, which is not known, and of the basic study design, which typically is conditioned by factors outside of the evaluator's control (e.g., the size and nature of the units of intervention and the properties of the outcomes of interest).

There are six prominent discretionary factors associated with statistical power of

particular study samples and sample size requirements to achieve a specified statistical power. One is the minimum relevant size impact, by which we mean the smallest size impact it is important to detect, if it exists. The second is the adopted level of statistical significance (α) or probability of making a Type I error (i.e., concluding there is an impact when there really is not). Commonly, evaluators set α equal to .05. A third discretionary factor is the desired level of statistical power (1- β), where β is the probability of making a Type II error (failing to detect a true impact if it occurs). Commonly, evaluators adopt a power level of .80. A fourth factor pertains to use of one-tailed or two-tailed testing, with two-tailed testing being most common. A fifth factor relates to use covariates to reduce measurement error (Bloom, 2006; Bloom, Richburg-Hayes & Rebeck Black, 2007), and a sixth factor relates to whether to assume fixed or random effects across sample blocks or clusters, which relates to the intended application of the study findings.

There are five especially notable inherent factors associated with the minimum detectable effect size or required sample size estimates associated with particular evaluation goals: (1) the size of the true average impact of the treatment or intervention (typically expressed in effect-size units); (2) for cluster (group) designs, the intra-class correlations (ICC) indicating the fraction of the total variance in outcome that lies between clusters; (3) the number of sample units within clusters; (4) the proportion of the sample expected to be in the treatment (or comparison) group (Bloom, 2006; Bloom et al., 2008; Hedges & Rhoads, 2010; Konstantopoulos, 2008a, 2008b; Raudenbush, 1997; Raudenbush, Martinez, & Spybrook, 2007; Schochet, 2008a); and (5) the minimum relevant effect size. For blocked random assignment design studies, the variability in impacts across blocks or effect size heterogeneity also affects the minimum detectable effect size and minimum required sample size (Raudenbush, Martinez, & Spybrook, 2007; Hedges &

Rhoads, 2010; Konstantopoulos, 2008b, 2009)[1].

For RD design studies, an inherent factor in determining minimum detectable effect size or minimum required sample size is the ratio of the asymptotic variances of impact estimators of RD design and experimental design, referred to as the "design effect." For single level RD design studies, the design effect can be expressed as $\frac{1}{1-\rho_{TS}^2}$, where ρ_{TS} is the correlation between treatment status and the criterion measure used to determine whether or not the unit was assigned to the treatment group (Schochet 2008b). Notably, ρ_{TS} will vary depending on three factors: (1) The distribution of the criterion measure in the population that is represented by the study sample; (2) the location of the cut-off score in this distribution; and (3) the proportion of the sample that is in the treatment group (Schochet, 2008b). The resulting consequence of the design effect for the statistical power of a particular study design is detailed in the Appendix and described in Schochet (2008b).

PowerUp! allows the user to compute either the minimum detectable effect size (MDES) or the minimum required sample size (MRSS) for studies by specifying inherent and discretionary factors, based on the best available information about them. For example, the user can specify assumed unconditional ICCs, drawing on resources such as Hedges and Hedberg (2007) and the average size of clusters, based on demographic data. S/he can then set values for discretionary factors, such as the desired level of statistical precision, the nature of statistical controls that will be used, and the relative size of the treatment and comparison groups. Within each design, the user may select other design features, including the number of levels of clustering or blocking, the nature of the cluster or block effect, and the expected level of sample attrition.

The minimum relevant effect size is both one of the most important factors and one that

requires considerable judgment on the part of the evaluator. It also is frequently is not explicitly discussed in evaluation design reports or considered in evaluating study findings.

SELECTING THE MINIMUM RELEVANT EFFECT SIZE (MRES)

Most often power analysis entails estimating and evaluating minimum detectable effect sizes (MDES) for specific sample sizes and designs. *PowerUp!* is designed to encourage and facilitate designing studies with adequate power to detect impacts equal to or larger than an established minimum size that has relevance for policy or practice. We refer to this as the minimum relevant effect size (MRES). In some cases, there is an empirical or policy basis for establishing a minimum size impact that is relevant and, thus, for a "target" MRES to use in designing a study or as the basis for judging the adequacy of an existing study sample to estimate reliably whether or not a treatment has a meaningful effect. The two obvious considerations in deciding on the MRES are cost and actual size of impact. For example, a costly educational intervention such as lowering class size would have practical relevance only if it generates relatively large impacts on student achievement, whereas a low-cost intervention such as financial aid counseling would need to have only modest impacts on college attendance for it the findings to have practical relevance. Alternatively, often there may be threshold effects that are needed before an intervention would be judged to be important for policy. For example, even a low-cost intervention that moves student achievement one or two points on a 500 point scale is not likely to have practical importance, regardless of whether or not the study findings are statistically significant.

Educators frequently describe their goals for changes in policy or practice in terms of their potential to close achievement gaps (e.g., between gender or race/ethnic groups) or in

relation to an average year of student growth in the outcome of interest. Importantly, these types benchmarks are sensitive to the metrics used (Bloom, Hill, Black, & Lipsey, 2008; Hill, Bloom, Black, & Lipsey, 2007). Thus, it is generally best to use natural units (like test score gains or percentage point reductions in dropout rates) for determining the minimum relevant size impact and, subsequently, convert this to effect size units (the MRES).

COMPUTING THE MINIMUM DETECTABLE EFFECT SIZE (MDES)

A convenient way to determine whether or not a completed study has adequate statistical power is to compute the MDES and compare this with the MRES. A priori, the goal is to design the study such that the MDES is less than or equal to the MRES and, thereby, maximize the chance that, if no impacts are detected, it is pretty certain that any true impacts escaping detection were sufficiently small as to have no practical or policy significance.

In contrast to the MRES, which is independent of the study design, the MDES depends on the actual sample design that was (or will be) implemented. Specifically, it is the minimum true effect-size that a particular study can detect with a specified level of statistical precision and power. The MDES depends on a variety of characteristics of the actual study including the study design, the extent and nature of clustering, the total sample size available for analysis (e.g., taking account of sample attrition), and the allocation of the sample to treatment and control conditions.

In general, the formula for estimating the MDES (in standard deviation units) can be expressed as:

$$MDES = M_v * SE / \sigma$$

where M_v is the sum of two *t*-statistics (Bloom, 1995, 2005, 2006; Murray, 1998). For one-

tailed tests, $M_v = t_\alpha + t_{1-\beta}$ with v degrees of freedom (v is a function of sample size and number of covariates) and for two-tailed tests (which are typically applied in studies designed to measure treatment effects), $M_v = t_{\alpha/2} + t_{1-\beta}$. *SE* is the standard error of the treatment effect estimate, and σ is the pooled total standard deviation of the outcome. (Throughout this paper and in the accompanying tools, the effect size has been defined as the difference in raw score units of the outcome of interest, divided by the pooled total standard deviation.)

Figure 1 below illustrates the construct of the multiplier for one-tailed tests. It is the distance in standard error (*t*-statistic) units such that, if the null hypothesis ($H_0 : \overline{Y}_T - \overline{Y}_C = 0$) is true, the Type I error is equal to α and, if the alternative hypothesis ($H_a : \overline{Y}_T - \overline{Y}_C > 0$) is true, the Type II error is equal to β. Put another way, the MDES is the smallest size true effect we expect to be able to detect with the specified power and precision.

----- Insert Figure 1 about here -----

It is possible to calculate the MDES for any study design as long as the ratio of the *SE* to σ is known and other key assumptions about the study design and analytic model have been specified (e.g., the sample size and its allocation across clusters and to treatment conditions, the level at which random assignment occurs, the number of covariates and their explanatory power, and the level of sample attrition). For example, in the two-level simple cluster random assignment design where treatment is at level 2 (Table 1, model 3.1), the treatment effect can be estimated using a 2-level hierarchical linear model:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2_{|X})$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02} W_j + \mu_{0j}, \mu_{0j} \sim N(0, \tau^2_{|W})$
$\beta_{1j} = \gamma_{10}$

Reduced form: $Y_{ij} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}W_j + \gamma_{10}X_{ij} + \mu_{0j} + r_{ij}$

In this case, the MDES formula from Bloom (2006, p.17) is as follows:[3]

$$MDES = M_{J-g^*-2}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}}$$

where,

Multiplier for one-tailed test: $M_{J-g^*-2} = t_{\alpha} + t_{1-\beta}$ with *J- g**-2 degrees of freedom;

Multiplier for two-tailed test: $M_{J-g^*-2} = t_{\alpha/2} + t_{1-\beta}$ with *J- g**-2 degrees of freedom;

J = the total number of clusters;

*g** = the number of group covariates used;

$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$ is the unconditional intra-class coefficient (ICC);

τ^2 = Level-2 (between group-level) variance in the unconditional model (without any covariates);

σ^2 = Level-1 (individual-level) variance in the unconditional model;

$R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$ is the proportion of variance in the outcome measure occurring at level one that is explained by covariates, *X*;

$R_2^2 = 1 - (\tau_{|W}^2 / \tau^2)$ is the proportion of the within group variance (at level two) that is explained by the covariates, *W*;

P = the proportion of this sample assigned to the treatment group (J_T / J);

Sample attrition reduces statistical power by lowering the size of the analytic sample[4]. For the 2-level simple cluster random assignment design, attrition might occur at both levels. Suppose the sample retention rates (=1- the percent of the study sample lost to follow up) at

levels 1 and 2 are r_1 and r_2, respectively, the revised *MDES* formula containing the retention rates is[5]:

$$MDES = M_{Jr_2-g^*-2}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)Jr_2}+\frac{(1-\rho)(1-R_1^2)}{P(1-P)Jnr_2r_1}}$$

In addition to Bloom's (1995, 2005, 2006) work on the background and computation of minimum detectable effect sizes, the specific formulae for *SEs* used in the *PowerUp!* tools rely on the work of others. For example, formulae for the 2-level simple cluster random assignment design studies (Table 1, model 3.1) draw on Raudenbush (1997); those for 2-level blocked individual random assignment designs (Table 1, models 2.2 and 2.3) draw on Raudenbush & Liu (2000); those for 3-level simple cluster random assignment designs (Table 1, model 3.2) and for 3-level blocked individual or cluster random assignment designs (Table 1, models 2.4, 3.1 and 3.2) draw on Hedges & Rhoads (2010), Konstantopoulos (2008a, 2008b and 2009b), Schochet (2008a), and Spybrook (2007); those for 4-level blocked cluster random assignment designs with treatment at level 3 (Table 1, models 4.4 and 4.5) draw on Spybrook (2007); those for the various regression discontinuity designs (Table 1, models 5.1-5.6) draw on Schochet (2008b); and those for the interrupted time-series designs (Table 1, model 6.0) draw on Bloom (1999, 2003). Notably, because the *SE* can be expressed in terms of the pooled standard deviation of the outcome, it also could be expressed in terms that are related to the *SE*, such as the unconditional ICC or the *R*-squared. The MDES formulae for 4-level simple cluster designs and the other 4-level blocked random assignment designs were derived following similar logic as applied for the above designs. The MDES formulae for all of the designs described above and listed in Table 1 above are presented in Appendix A.

These MDES formulas are the basis for the *PowerUp!* tools in the accompanying

Microsoft Excel™ workbook.

COMPUTING MINIMUM REQUIRED SAMPLE SIZES

The same formulae that are used to compute minimum detectable effect sizes can be manipulated to work in reverse to determine the minimum size sample required to ensure that the MDES for a study will be less than or equal to the minimum size that is relevant for policy or practice (the MRES). For example, using the MDES formula for a 2-level cluster random assignment design study in the example above (Table 1, model 3.1), the sample size (J) can be expressed as follows:

$$J = \left(\frac{M_{Jr_2 - g^* - 2}}{MDES} \right)^2 \left(\frac{\rho(1 - R_2^2)}{P(1-P)r_2} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)nr_2r_1} \right)$$

Because the multiplier, $M_{Jr_2 - g^* - 2}$, is a function of *J*, *PowerUp!* solves for the sample size through an iterative process that is illustrated below.

USING THE *PowerUp!* TOOLS

PowerUp! users have options to compute a minimum detectable effect size (MDES) or to determine the minimum required sample size (MRES) by selecting the core study design features from among the 21 identified in Table 1. *PowerUp!* is a user-friendly, flexible tool that complements existing power analysis software by answering one of two questions, based on user-supplied assumptions: (1) What is the minimum size of true impact that will have a specified likelihood of being detected (i.e., the MDES), given a specified study design, sample size, and allocation to treatment and control condition? and (2) for a given study design, what is the minimum required sample size (MRSS) to have the desired power to detect a true impact (if

such exists) that is at or above a minimum size that is relevant for policy or practice (i.e., the MRES)?

Computing Minimum Detectable Effect Sizes (MDES**)**

In general, one would want to estimate the minimum detectable effect size (MDES) in cases where there is an existing sample design—for example, to determine whether a completed study used a sufficiently large sample to expect that it would have detected impacts of a given size or larger, if they occurred. Put another way, one could determine whether the study was likely to have uncovered true impacts that were equal to or larger than a specified minimum relevant effect size (MRES).

The first step in using the tool is to determine the basic study design, using Table 1 above and then select the corresponding Tab in the *PowerUp!* tool. For example, a study that used simple individual random assignment of students to treatment or control condition corresponds to Model 1.0 (IRA) in Table 1. Thus, the user would to select Tab 1.0 IRA to enter the tool for computing the study's MDES (see Table 3). Once in the relevant worksheet, the user has the opportunity to supply relevant assumptions and preferences in the highlighted cells. These include the desired statistical precision and power (i.e., α, whether a one- or two-tailed test is being used, and 1-β); assumptions about the analytic models that were used to estimate impacts (i.e., the proportion of variance in the outcome explained by the model and the number of covariates used); and characteristics of the study sample (i.e., total size and proportion assigned to treatment condition).

Table 3 shows input and output for computing the MDES for a study that randomly assigned 240 individuals to treatment or control condition in equal proportion. The user in this example specified an alpha level of .05, a two-tailed test, 80 percent power, and an R^2 of .60.

After inputting these user-supplied assumptions, *PowerUp!* returned an estimate of the MDES equal to .230 (shown in bright green at the bottom of the worksheet).

--- Insert Table 3 about here---

Estimating the MDES for a blocked individual random assignment design study is similar, only the user would select a Tab corresponding to the relevant version of model 2, depending on whether there are 2, 3 or 4 levels of clustering and whether the analytic model assumes constant, fixed, or random effects, respectively. Table 4 illustrates the worksheet for model 2.3, which has 2 levels of clustering and assumes random effects. In general, the tools for blocked random assignment designs work similarly to those above, with the addition of treatment effect heterogeneity parameters (ω) that denote assumptions about the variability of the treatment effect across certain level of blocks, standardized by the outcome variation at that level of block (Hedges & Rhoads, 2010; Konstantopoulos, 2008b, 2009). Furthermore, the block-level variance explained is the proportion of variance between blocks on the treatment effect that is explained by block-level covariates, not the block-mean variance explained.

----- Insert Table 4 about here -----

An example of studies for which it would be appropriate to use these block random assignment designs is for the recently completed study of charter schools (Gleason, Clark, Tuttle, & Dwoyer, 2010), where lotteries were used to assign eligible charter school applicants for over-subscribed schools to admission (treatment group) or not (control group) for the particular charter school to which they applied (the block). In analyzing the data from this randomized block design study, it was assumed that the effects were unique to each charter school.

For a cluster random assignment design, the user would select the relevant Model 3 or Model 4 tab from the *PowerUP!* tools. Models 3.1 through 3.3 all pertain to simple cluster

random assignment designs, but differ in the number of levels of clustering. For example, a study in which treatment occurred at the cluster level closest to the unit of analysis (e.g., randomization occurred at the classroom level and students were the unit of analysis), the user should select model 2.2 (CRA2_2r). If schools were randomized to treatment or control status and students clustered in classrooms were the unit of analysis, the user should select model 3.2; and, if districts were randomized to treatment or control conditions, but students clustered in schools and classrooms were the unit of analysis, the user should select model 3.3.

In addition to the input required for individual random assignment design studies, for cluster random assignment designs, the user also needs to provide information on the unconditional intra-class correlations between the analysis units and the clusters, assumptions about the proportion of variance in the outcome explained by the estimation model for each level of the data, and details about the size of sample clusters as well as the overall sample. The yellow cells in Table 5 illustrate these input parameters for a design with four levels of clustering and where the treatment occurs at the highest level of clustering (Table 1, model 3.3, and *PowerUp!* Tab 3.3 CRA4_4r).

----- Insert Table 5 about here -----

In this particular example, the user selected the same basic parameters as in the previous example shown in Table 3 (e.g., alpha level, two-tailed testing, power standard, and proportion assigned to treatment condition). However, in this case, the user also needed to provide assumptions or actual data about the intra-class correlations (ICCs) at the various levels, the proportion of variance in the outcomes explained by covariates (R_j^2) at the various levels, the number of sample units randomized, and the average number of units in each of the clusters. In this particular example, there are 1200 analysis units, clustered as follows: 10 level-1 units per

level 2 cluster; 2 level-2 units per level 3 cluster; 3 level-3 units per level 4 cluster; and 20 level-4 units. The user specified that covariates will be included in the levels 1-4 analyses, and that covariates will explain half of the variance in the outcome measured at each level. With only one covariate in the level four analysis (which is the level at which randomization was conducted), the MDES is estimated to be .292 standard deviations (shown in the green box).

For blocked cluster random assignment designs, the user selects one of the model 4 tabs from the *PowerUp!* tools. As with the blocked individual random assignment models, the appropriate tab depends on the level of blocking at which random assignment occurred and whether the analysis is designed to estimate fixed or random effects.

As discussed above, the *PowerUp!* tool uses formulas for computing the MDES and minimum sample size requirements for regression discontinuity designs (RD) that are "derivative" of those used for random assignment designs. Essentially, in an RD design, the treatment and control groups are not determined by randomization, but by a specific rule for sorting based on a continuous criterion variable. As a result, for any given set of study design parameters (e.g., alpha level, one or two tailed test, power level, and MRES), the MDES and MRSS are considerably larger than under an individual random assignment design. *PowerUP* has built into the MDES formulas estimates of the "penalty" (formally referred to as the "design effect") based on work by Schochet (2008b).

The design effect can be thought of as the multiplier on the sample size requirement needed for the RD design to have similar statistical power to a block random assignment design study. For example, Schochet (2008b) found that, for a study in which the optimal functional form is linear, the criterion measure is normally distributed, the cutoff score is at the mean, and 50 percent of the sample is in the treatment group, the estimated design effect is 2.75. Drawing

on this work, *PowerUp*! includes this value as the default design effect in the RD design tools. Importantly, Schochet (2008b, p. 10) notes that "T(t)he linearity (and constant treatment effects) assumptions will likely provide a lower bound on RD design effects." Thus, *PowerUp!* users may want to modify this default.

Table 6 illustrates the *PowerUp!* tool for computing the MDES for model 5.2 (*PowerUp!* Tab 5.2 RD2_1r), which would apply to a study in which students were the units of assignment and school effects are random. This model is analogous to 2-level random effect block random assignment designs (Table 1, model 2.3 and *PowerUp!* Tab 2.3 BIRA2_1r). In addition to the parameters required to calculate MDES for 2-level random effect blocked random assignment design study (e.g., the treatment effect heterogeneity, ω , and the proportion of variance between blocks on the treatment effect explained by the block-level covariates), the user also needs to accept or change the default design effect.

----- Insert Table 6 about here -----

For interrupted time-series design (ITS) studies, the user needs to specify four design parameters that are unique to the ITS design: (1) the number of baseline years of data; (2) the follow-up year of interest; and (3) whether an additional comparison group is used; and (4) if an additional comparison group is used, its size relative to that of the number of treatment group units.

Table 7 illustrates the MDES calculation for an ITS design study with an alpha level of .05, using a 2-tailed test, and 80 percent power. The ICC for the cohorts is 0.02, with 5 waves of baseline data, 6 program schools, and 200 students per school. The proportion of variance of between-cluster (cohort) explained by a cohort-level covariate is 0.2. The impact is estimated in the second observation period following the treatment. For this example, the MDES is estimated

to be 0.20 assuming the study sample does not include any no-treatment comparison units and 0.24 if two-thirds of the sample consists of non-treatment comparison units. The reason the MDES increases if some of the sample comes from no-treatment comparison units is that, holding sample size constant, the no-treatment comparison units increase the standard error of the impact estimate increases due to the need to make additional comparisons.

----- Insert Table 7 about here -----

Determining the Minimum Required Sample Size to Achieve Study Goals

The logical way to design a study is to begin by determining the most appropriate, feasible study design. The first step is to determine whether it is feasible to conduct an experimental design evaluation. The second step is to determine the appropriate units for assignment to treatment condition and for analysis, considering factors like the natural unit for delivering the intervention, the potential for contamination of the control group, and the likelihood of gaining necessary cooperation from partner organizations and individuals. The third step is to determine the minimum relevant effect size (see discussion above), considering factors such as the cost of the intervention and the nature of the target outcome (e.g., student test scores, or high school graduation rates), their means and standard deviations for target populations, and "normal" changes in the levels over time. The fourth step is to use this information in estimating how large the study sample needs to be to achieve the study objectives.

Typically, design teams arrive at their target sample sizes in one of two ways. One way is to figure out how large a sample can be supported by the evaluation dollars, determine the MDES implied by that sample size, and rationalize it. Another common strategy is to use a trial and error approach to computing the MDES for various sample sizes until converging on an "acceptable" MDES. The *PowerUp!* sample size estimation tools use variants of the formulae

used to calculate minimum detectable effect sizes to compute sample size requirements for user-defined minimum detectable effect sizes (which should be the same or smaller than the minimum relevant effect size). It does this through an iterative process, which we have automated through an Excel macro, which works as follows:

- Step 1. An initial "guesstimate" of the sample size (individuals, clusters or blocks) has been set at 30.
- Step 2. An estimate of the multiplier and the minimum required sample size is calculated using the formulas based on the "guesstimate" of the sample size.
- Step 3. If the "guestimate" of the sample size differs from that calculated using the formula, the "guestimate" is replaced with the average of the original "guestimate" and the calculated sample size, and the program goes back to Step 2.
- The process stops when the difference between the calculated sample size and "guestimate" is within ± 0.1.

PowerUp! includes a tab for computing minimum required sample sizes to achieve a specified MRES under user-defined parameters for each of the 21 study design configurations specified in Table 1. The example shown in Table 8 pertains to a simple 2-level cluster random assignment design study where the analysis will be conducted using a 2-level hierarchical linear model with student nested within schools and assuming random effects. In this example, the user's desired MDES is 0.25 standard deviations. The user also has specified an alpha of .05, two-tailed tests of statistical significance, and 80 percent power level. The user has assumed an ICC of 0.20, that there will be an average of 60 students per school, and that 90 percent of the schools and 80 percent of the students in the original study sample will be retained in the analysis. The analysis will include one student-level covariate that explains 50 percent of

student-level variance, and one school-level covariate explaining 70 percent of school-level variance. Using these user-supplied parameters, *PowerUp!* estimates that the study sample should include 41 schools (2,460 students).

----- Insert Table 8 about here -----

Users may find the PowerUP! tools useful for exploring the implications for sample size requirements of varying the study design parameters. For example, it would be relatively easy to assess the sensitivity of the MRSS to the ICC level, to decisions about blocking, or to a higher R^2 that would result from investing in good pretest measures.

CONCLUSION

PowerUp! is intended as a complement to, not a replacement for, other sample design tools. One goal in developing this tool is to encourage and enable evaluators who are planning new evaluations to more accurately estimate the size samples needed to ensure a high probability that the study will detect meaningful size impacts, should they result from the intervention. A second goal is to make it easier for those who are judging the findings from existing research to estimate how large an impact would need to be in order for the study to have a reasonable chance of observing a statistically significant difference between the treatment and control groups. A third goal is make it easy for evaluators to see how sensitive evaluation findings may be to factors that are inherent to the particular intervention and setting for the study (e.g., the units for delivering the intervention, the age and demographics of the study sample, important outcomes), as well as discretionary factors (e.g., study design, levels of statistical power and precision required, fixed or random effects, control variables).

PowerUp! makes key assumptions study designs transparent. It also invites users to examine the sensitivity of their estimated sample size requirements or minimum detectable effect

sizes to various assumptions and decisions of the evaluator. For example, it would be easy for an evaluator to determine how much more sample would be required if the goal were to generate a population estimate of the impact rather than simply a sample estimate; determine the expected decrease in the MDES or the minimum required sample size if sample attrition could be reduced by 10 percentage points; estimate the "cost" of blocking as opposed to using a simple random assignment design; determine how sensitive the MDES or minimum sample size requirement is to the assumed ICC; and estimate the difference in the MDES for a give sample size if the evaluator opts to use an experimental design rather than a regression discontinuity or interrupted time series design.

PowerUp! makes it easy for the user to explore the implications of sample design decisions and user-supplied assumptions about unknowns, such as the ICC, explanatory power of covariates, and ultimate sample attrition rates. By using Microsoft EXCEL™ as the platform for this tool, we have made it possible for others to not only use the tool as we have designed it, but to adapt and enhance it to meet other objectives. .

REFERENCES

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics 114*: 533-575.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19* (5), 547-556

Bloom, H. S. (1999). Estimating program impacts on student achievement using "short" interrupted time series. New York, NY: MDRC.

Bloom, H. S. (2001). *Measuring the impacts of whole-school reforms: Methodological lessons from an evaluation of Accelerated Schools. New York, NY: Manpower Demonstration Research Corporation*. Accessed November 25, 2012 at http://www.mdrc.org/sites/default/files/full_439.pdf.

Bloom, H. S. (2003). Using short interrupted time-series analysis to measure the impacts of whole school reforms: With applications to a study of accelerated schools. *Evaluation Review*, 27(1): 3 – 49.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In Howard S. Bloom (editor), *Learning more from social experiments: Evolving analytic approaches*, 115-172, New York: Russell Sage Foundation.

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. MDRC Working papers on research methodology. Available online at: http://www.mdrc.org/publications/437/full.pdf

Bloom, H. S. (2009). *Modern regression discontinuity analysis*, MDRC: New York. http://www.mdrc.org/publications/539/abstract.html

Bloom, H. S., Richburg- Hayes, L. & 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), pp. 30–59.

Bloom, H. S., Hill, C. J., Black, A. R. & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1, pp. 289 – 328.

Borman, G., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N.A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal. 44*(3), pp. 701 –731.

Boruch, R.F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Cook, T. D., Hunt, H. D., & Murphy, R. F. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal, 37*,535-597.

Cook, T. D. & Wong, V. C. (2007). *Empirical tests of the validity of regression discontinuity designs.* Chicago, IL: Northwestern University.

Gamse, B. C., Bloom, H. S., Kemple, J. J., Jacob, R. T. et al. (2008). *Reading First impact study: Interim report* (NCES 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Educational Sciences, U.S. Department of Education.

Gleason, P., Clark, M., Tuttle, C., & Dwoyer, E. (2010). The evaluation of charter school impacts: Final report. NCEE 2101-4010. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences.

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-

randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1): 60–87.

Hedges, L. V. & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/.

Hill, C. J., Bloom, H.S. Black, A. R., & Lipsey, M.W. (2007). Empirical Benchmarks for Interpreting Effect Sizes in Educational Research. New York: MDRC Working Papers on Research Methodology. Retrieved October 21, 2011: http://www.mdrc.org/publications/459/full.pdf

Imbens, G. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142* (2), 615-635.

Jacob, B. & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics* 86, 1: 226-244

Kemple, J. J., Herlihy, C. M., & Smith, T. J. (2005). *Making progress toward graduation: Evidence from the Talent Development High School Model*. New York: Manpower Demonstration Research Corporation.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*, 66-88.

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*, 265-288.

Konstantopoulos, S. (2009). Using power tables to compute statistical power in multilevel

experimental designs. *Practical Assessment, Research & Evaluation*, 14(10). Retrieved April 21, 2010: : http://pareonline.net/getvn.asp?v=14&n=10.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics* 142: 829-850.

Murnane, R. & Willett, J. (2010). *Methods matter: Improving causal inference in educational and social science research.* Oxford University Press.

Murray, D. (1998). *Design and analysis of group-randomized trials*. Oxford: Oxford University Press.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class-size experiment. *Educational Evaluation and Policy Analysis* 21:127–42.

Orr, L. (1998). *Social experiments evaluating public programs with experimental methods*. Sage Publications: London.

Quint, J. C., Sepanik, S., & Smith, J. K. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessment of Student Thinking in Reading (FAST-R) program in Boston elementary schools*. New York: Manpower Demonstration Research Corporation.

Quint, J. C., Bloom, H. S., Rebeck Black, A. & Stephens, L., with Akey, T.M. (2005). *The challenge of scaling up educational reform: findings and lessons from First Things First.* New York: Manpower Development Research Corporation.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S. W. & Liu, X. (2000). Statistical power and optimal design for multisite

randomized trials. *Psychological Methods*, 5(2), 199-213.

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

Spybrook, J. (2007). *Examining the experimental designs and statistical power of group randomized trials funded by the Institute of Education Sciences.* Ann Arbor, MI: University of Michigan. (Unpublished Dissertation).

Spybrook, J., Raudenbush, S.W., Congdon, R. & Martinez, A. (2009). *Optimal design version 2.0.*

Schochet, P. Z. (2008a). Statistical power for randomized assignment evaluation of education programs. *Journal of Educational and Behavioral Statistics*. *33* (1), 62-87.

Schochet, P. Z. (2008b). *Technical methods report*: *Statistical power for regression discontinuity designs in education evaluations* (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., Smith, J. (2010). *Standards for regression discontinuity designs*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston; Houghton Mifflin.

Thistelwaite, D. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51* (6), 309-317.

Trochim, W. (1984). *Research design for program evaluation: The regression-discontinuity approach.* New York: Sage Publications.

U.S. Department of Education, What Works Clearinghouse (2008). *Procedures and standards handbook (version 2),* Washington: DC, December.

FOOTNOTES

[1]The authors contributed equally to this work. The paper and accompanying analytic tool benefitted greatly from input at various stages of the work by Spyros Konstantopoulos, Mark Lipsey, Larry Orr, Mike Weiss, and Kerry Hofer. We also thank two anonymous reviewers for their thoughtful comments and suggestions. The *PowerUp!* tool that accompanies this paper may be accessed at: http://peabody.vanderbilt.edu/research/pri/methods_resources.php or from the corresponding author.

[2]The effect size variability and effect size heterogeneity have different definitions but both indicate the variability/heterogeneity of treatment effect vary across block.

[3]When there is only one level-2 covariate, *W*, the variance of the main effect of treatment is estimated by (Raudenbush, 1997) as follows:

$$\text{var}(\hat{\gamma}_{01} \mid W) = \frac{4(\tau^2_{|w} + \sigma^2 / n)}{J}\left[1 + \frac{1}{J-4}\right],$$

where, $\hat{\gamma}_{01}$ is the point estimate of treatment effect; *W* is a level-2 covariate, and $\text{var}(\hat{\gamma}_{01} \mid W)$ represents the variance of treatment effect estimate after adjusting covariate, *W*; $\tau^2_{|w}$ is the level-2 variance after adjusting covariate, and σ^2 is the level-1 variance; *J* is the total number of level-2 units (clusters) and *n* is the sample size per cluster. The MDES can be expressed as:

$$MDES = M_{J-g^*-2}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}}\sqrt{1 + \frac{1}{J-4}}$$

The MDES formula derived by Bloom (2006) ignores the factor $\sqrt{1 + \frac{1}{J-4}}$. The reason is that this adjustment has no practical significance for very small samples (i.e., samples less than 30). For example, for *J* = 6, 10, 14, 20, and 30, respectively, the MDES calculated from Bloom's

formula is 18%, 7%, 5%, 3%, and 2% smaller than what would have been estimated including the small sample correction. Similarly, the standard error formulas for three-level hierarchical randomized assignment designs derived by other researchers did not include such as factor (e.g., Hedges & Rhoads, 2010; Konstantopoulos, 2008a; Spybrook, 2007; Schochet, 2008a), to be consistent we use the MDES formula by ignoring this factor.

[4]In addition to affecting the MDES and MRSS, sample attrition also may introduce the threat of bias (U.S. Department of Education, What Works Clearinghouse, 2008). *PowerUp!* does not address the threat of bias due to sample attrition.

[5]An alternative approach to handling attrition in power analysis is to calculate the required sample size to meet MDES when there is no attrition and then adjust this sample size by dividing (1- attrition rate).

Table 1: Study Designs and Analyses Included in the *PowerUp!* Tools

Study Design	1 Model Number	2 Simple or Blocked (Stratified) Assignment	3 Levels of Clustering	4 Unit of Treatment Assignment	5 Treatment Assignment Level	6 Cluster/ Block Effect	7 Worksheet Name for: MDES Calculation	8 Worksheet Name for: Sample Size Calculation
Experimental Designs								
Individual Random Assignment Designs (Level of Assignment = Level of Analysis)								
1. Simple Individual Random Assignment (IRA)	1.0	simple	1	individual	1	N/A	IRA	N_IRA
2. Blocked (Stratified) Individual Random Assignment (BIRA)	2.1	blocked	2	individual	1	constant	BIRA2_1c	N_BIRA2_1c
	2.2		2	individual	1	fixed	BIRA2_1f	N_BIRA2_1f
	2.3		2	individual	1	random	BIRA2_1r	N_BIRA2_1r
	2.4		3	individual	1	random	BIRA3_1r	N_BIRA3_1r
	2.5		4	individual	1	random	BIRA4_1r	N_BIRA4_1r
Cluster Random Assignment Designs (Level of Assignment ≠ Level of Analysis)								
3. Simple Cluster Random Assignment (CRA)	3.1	simple	2	cluster	2	random	CRA2_2r	N_CRA2_2r
	3.2		3		3	random	CRA3_3r	N_CRA3_3r
	3.3		4		4	random	CRA4_4r	N_CRA4_4r
4. Blocked (Stratified) Cluster Randomized Assignment Designs (BCRA)	4.1	blocked	3	cluster	2	fixed	BCRA3_2f	N_BCRA3_2f
	4.2	blocked	3	cluster	2	random	BCRA3_2r	N_BCRA3_2r
	4.3	blocked	4	cluster	2	random	BCRA4_2r	N_BCRA4_2r
	4.4	blocked	4	cluster	3	fixed	BCRA4_3f	N_BCRA4_3f
	4.5	blocked	4	cluster	3	random	BCRA4_3r	N_BCRA4_3r
Quasi-experimental Designs								
5. Regression Discontinuity Designs (RD)	5.1	blocked	2	individual	1	fixed	RD2_1f	N_RD2_1f
	5.2	blocked	2	individual	1	random	RD2_1r	N_RD2_1r
	5.3	simple	2	cluster	2	random	RDC_2r	N_RDC_2r
	5.4	simple	3	cluster	3	random	RDC_3r	N_RDC_3r
	5.5	blocked	3	cluster	2	fixed	RD3_2f	N_RD3_2f
	5.6	blocked	3	cluster	2	random	RD3_2r	N_RD3_2r
6. Interrupted Time-Series Designs (ITS)	6.0	blocked	3	cluster	2	constant at level 3; random at level 2	ITS	N_ITS

Table 2: Examples of Blocked Random Assignment Designs

	Levels of Blocking									
	Two (students and schools)			Three (students, schools, and districts)			Four (students, schools, districts and states)			
PowerUP! Model →	BIRA2_1c	BIRA2_1f	BIRA2_1r	BIRA3_1r	BCRA3_2f	BCRA3_2r	BIRA4_1r	BCRA4_2r	BCRA4_3f	BCRA4_3r
Level of Random Assignment	1 = Students			1 = Students	2 = Schools		1 = Students	2 = Schools	3 = Districts	
Block Effects	c	f	r	r	f	r	r	r	f	r

Note: c = Constant block effects model; f = Fixed block effects model; r = Random block effects model.

Table 3: Sample Tool for Computing the MDES for a Simple Individual Random Assignment Design Study (See Table 1, Model 1.0 and *PowerUP!* Tab 1.0 IRA)

Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
P	0.50	Proportion of the sample randomized to treatment: $n_T / (n_T + n_C)$
R^2	0.60	Percent of variance in outcome explained by covariates (See Bloom et al. 2007; Deke et al. 2010)
k*	1	Number of covariates used
n (Total Sample Size)	240	
M (Multiplier)	2.81	Computed from T_1 and T_2
T_1 (Precision)	1.97	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.84	Determined from given power level
MDES	**0.230**	Minimum Detectable Effect Size

START OVER

Note: The parameters in the yellow cells need to be specified. The MDES will be calculated automatically.

References:

Bloom, H. S., Richburg- Hayes, L. & 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), pp. 30–59.

Deke, John, Dragoset, Lisa, and Moore, Ravaris (2010). Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/pubs/20104003/

Table 4: Sample Tool for Computing the MDES for a 2-Level Blocked Individual Random Assignment Design Study with Random Effects (See Table 1, Model 2.3 and *PowerUp!* Tab 2.3 BIRA2_1r)

Assumptions		Comments
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.35	Proportion of variance in outcome between clusters (See Hedges and Hedberg 2007)
ω	0.10	Treatment effect heterogeneity: variability in treatment effects across Level 2 units, standardized by the variability in the Level-2 outcome
P	0.50	Proportion of Level 1 units randomized to treatment: $n_T / (n_T + n_C)$
R_1^2	0.00	Proportion of variance in the Level 1 outcome explained by Block and Level 1 covariates (See Bloom et al. 2007; Deke et al. 2010)
R_{2T}^2	0.00	Proportion of between block variance in treatment effect explained by Level 2 covariates (See Deke et al. 2010)
g*	0	Number of Level 2 covariates
n (Average Block Size)	80	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
J (Sample Size [# of Blocks])	480	Number of Level 2 units in the sample
M (Multiplier)	2.81	Computed from T_1 and T_2
T_1 (Precision)	1.96	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.84	Determined from given power level
MDES	**0.033**	Minimum Detectable Effect Size

START OVER

Note: The parameters in yellow cells need to be specified. The MDES will be calculated automatically.

References:

Bloom, H. S., Richburg- Hayes, L. & 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), pp. 30–59.

Deke, John, Dragoset, Lisa, and Moore, Ravaris (2010). Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/pubs/20104003/

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29(1): 60–87.

Table 5: Sample Tool for Computing the MDES for a 4-Level Simple Cluster Random Assignment Design Study with Treatment at Level 4 (See Table 1, Model 3.3 and *PowerUp!* Tab 3.3 CRA4_4r)

Assumptions		Comments
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho_4 (ICC_4)	0.05	Proportion of variance among Level 4 units (V4/(V1 + V2 + V3 + V4)) (See Hedges and Hedberg 2007)
Rho_3 (ICC_3)	0.05	Proportion of variance among Level 3 units (V3/(V1 + V2 + V3)) (See Hedges and Hedberg 2007)
Rho_2 (ICC_2)	0.10	Proportion of variance among Level 2 units (V2/(V1 + V2 + V3 + V4)) (See Hedges and Hedberg 2007)
P	0.50	Proportion of Level 4 units randomized to treatment
R_1^2	0.50	Proportion of explained variance in the Level 1 outcome by Level 1 covariates (See Bloom et al. 2007; Deke et al. 2010)
R_2^2	0.50	Proportion of variance in the Level 2 mean outcome explained by Level 2 covariates (See Bloom et al. 2007; Deke et al. 2010)
R_3^2	0.50	Proportion of variance in the Level 3 mean outcome explained by Level 3 covariates (See Bloom et al. 2007; Deke et al. 2010)
R_4^2	0.50	Proportion of variance in the Level 4 mean outcome explained by Level 4 covariates (See Bloom et al. 2007; Deke et al. 2010)
g_4*	1	Number of Level 4 covariates
n (Average Sample Size for Level 1)	10	Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)
J (Average Sample Size for Level 2)	2	Mean number of Level 2 units per Level 3 unit (harmonic mean recommended)
K (Average Sample Size for Level 3)	3	Mean number of Level 3 units per Level 4 unit (harmonic mean recommended)
L (Sample Size [# of Level 4 units])	20	Number of Level 4 units
M (Multiplier)	2.97	Computed from T_1 and T_2
T_1 (Precision)	2.11	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.86	Determined from given power level
MDES	**0.292**	Minimum Detectable Effect Size

START OVER

Note: The parameters in yellow cells need to be specified. The MDES will be calculated automatically.

References:

Bloom, H. S., Richburg- Hayes, L. & 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), pp. 30–59.

Deke, John, Dragoset, Lisa, and Moore, Ravaris (2010). Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/pubs/20104003/

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29(1): 60–87.

Table 6: Sample Tool for Computing the MDES for RD Design Analogous to 2-Level Blocked Individual Random Assignment Design with Random Effects (See Table 1, model 5.2 and *PowerUp!* Tab 5.2 RD2_1r)

Assumptions		Comments
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.15	Proportion of variance between clusters (Hedges & Hedberg 2007)
ω	0.20	Treatment effect heterogeneity or variance in treatment effect across Level 2 units, standardized by the Level-2 outcome variation: $\omega = \tau^2_{T2} / \tau^2_2$
P	0.50	Proportion of individuals randomized to treatment: $n_T / (n_T + n_C)$
R_1^2	0.50	Proportion of variance in the Level 1 outcome explained by the Level 1 covariates (See Bloom et al. 2007; Deke et al. 2010)
R_{2T}^2	0.10	Proportion of variance in treatment effect between Level-2 blocks explained by Level-2 covariates
g*	1	Number of Level 2 covariates
n (Average Cluster Size)	20	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
J (Sample Size [# of Clusters])	40	Number of Level 2 units in sample
Design Effect	2.75	Estimated from empirical data (last two rows) or based on other assumptions (Schochet, 2008)
M (Multiplier)	2.88	Computed from T_1 and T_2
T_1 (Precision)	2.02	Computed from given alpha Level, two-tailed or one-tailed test
T_2 (Power)	0.85	Computed from given power level
MDES	**0.232**	Minimum Detectable Effect Size
ρ_{TS}	0.8	Correlation between TREATMENT indicator and the score used for treatment assignment
Estimated Design Effect	2.78	Estimated multiplier on sample size for Randomized Block Design study with equal power (See Schochet 2008)

BACK TO RDD INTERFACE | START OVER

Note: The parameters in yellow cells need to be specified. The MDES will be calculated automatically.

References:

Bloom, H. S., Richburg- Hayes, L. & 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), pp. 30–59.

Deke, John, Dragoset, Lisa, and Moore, Ravaris (2010). Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/pubs/20104003/

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29(1): 60–87.

Schochet, P. Z. (2008). Technical methods report: Statistical power for regression discontinuity designs in education evaluations (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Table 7: Sample Tool for Computing the MDES for an Interrupted Time Series (ITS) Design Study: 3-Level Blocked Design with Random Effects at Level 2 and Constant Effects at Level 3 (See Table 1, Model 6.0 and *PowerUP!* Tab ITS)

Assumptions		Comments
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.02	Proportion of variance between cohorts (See Hedges and Hedberg 2007)
T (the number of baseline years)	5	The number of years prior to intervention for which the baseline, or pre-intervention, trend is established.
n (Average Sample Size for Level 1)	200	Mean number of Level 1 units per Level 2 unit, or cohort (harmonic mean rec
m (Sample Size [# of program schools])	6	The number of Level 3 units in the sample
R_2^2	0.20	Percent of variance in the outcome explained by covariates at Level 2
t_f (follow-up year of interest)	2	Year in which the outcomes are to be compared (i.e., "0" would indicate the year that treatment occurs; "1" would indicate the first year following the
g*	1	Number of Level 2 (cohort-level) covariates
Ratio of comparison units to experimental units	2	(# comparison schools / # program schools) at block level
M (Multiplier)	2.90	Computed from T1 and T2. 2.5 was used in Bloom (1999).
T_1 (Precision)	2.05	Computed from given alpha Level, two-tailed or one-tailed test. df=m*T-g-1
T_2 (Power)	0.85	Computed from given power Level. df=df=m*T-g-1
MDES (no comparison units)	**0.20**	Minimum Detectable Effect Size
MDES (with comparison units)	**0.24**	Minimum Detectable Effect Size

START OVER

Note: The parameters in yellow cells need to be specified. The MDES will be calculated automatically. This calculation assumes a design in which individuals are nested within successive grade cohorts in a school; cohort is a random effect; the school is constant effect.

References:

Bloom, H. S. (1999). Estimating program impacts on student achievement using "short" interrupted time series. New York, NY: MDRC.

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29(1): 60–87.

Table 8: A Sample Tool for Computing the Minimum Required Sample Size for Simple Two-Level Cluster Random Assignment Design with Treatment Occurring at Level 2 (See Table 1, Model 3.1 and *PowerUp!* Tab N_CRA2_2r)

Assumptions		Comments
MRES = MDES	0.25	MRES = MDES
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.20	Proportion of variance in outcome that is between clusters (See Hedges and Hedberg 2007)
n (Average Cluster Size)	60	Mean number of Level 1 units per Level 2 cluster (harmonic mean
Sample Retention Rate: Level 2 units	90%	Proportion of Level 2 units retained in analysis sample
Sample Retention Rate: Level 1 units	80%	Proportion of Level 1 units retained in analysis sample
P	0.500	Proportion of sample randomized to treatment: J_T / (J_T + J_C)
R_1^2	0.500	Proportion of variance in Level 1 outcome explained by Block and Level 1 covariates (See Bloom et al. 2007; Deke et al. 2010)
R_2^2	0.700	Proportion of variance in Level 2 outcome explained by Block and Level 2 covariates (See Bloom et al. 2007; Deke et al. 2010)
g*	1	Number of Level 2 covariates
M (Multiplier)	2.89	Computed from T_1 and T_2
J (Sample Size [Clusters #])	41	Number of clusters needed for given MRES
	RUN	START OVER

Note: The parameters in yellow cells need to be specified. Then click "RUN" to calculate sample size.

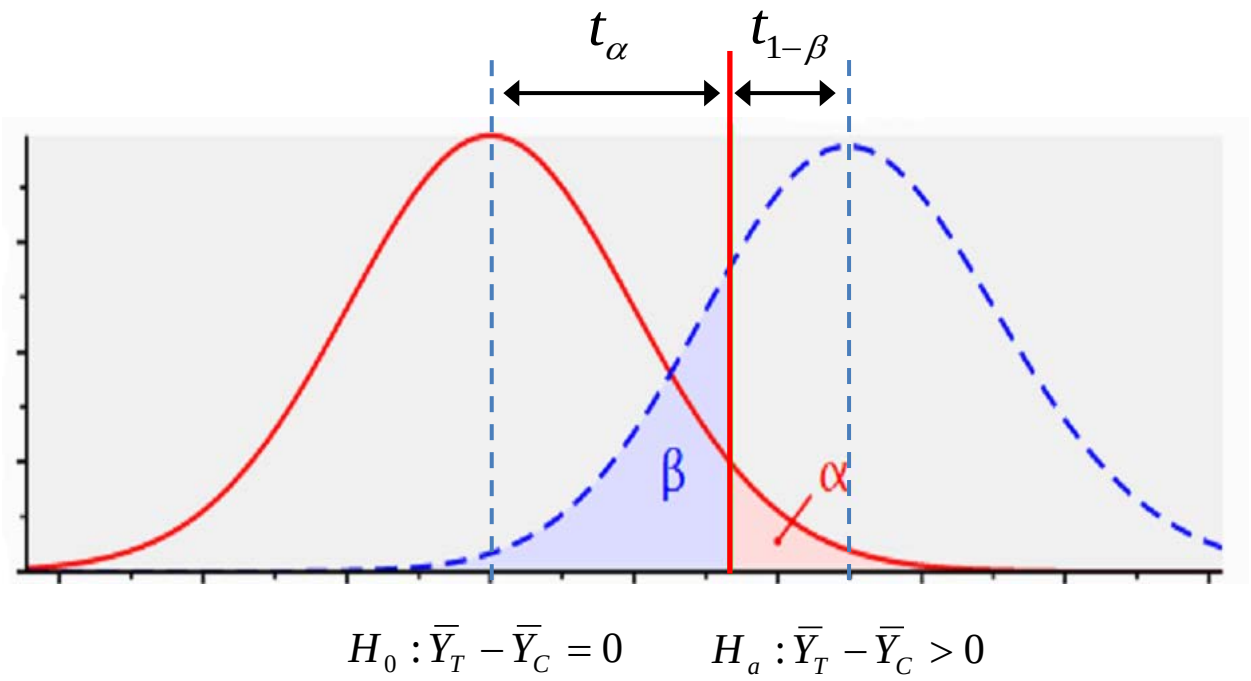
References:

Bloom, H. S., Richburg- Hayes, L. & 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), pp. 30–59.

Deke, John, Dragoset, Lisa, and Moore, Ravaris (2010). Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials (NCEE 2010-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/pubs/20104003/

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29(1): 60–87.

Figure 1: One-tailed Multiplier ($M_v = t_\alpha + t_{1-\beta}$)



Note: Adapted from Bloom (2006, Figure 1, page 22). The two-tailed multiplier: $M_v = t_{\alpha/2} + t_{1-\beta}$

APPENDIX: STATISTICAL MODELS, MINIMUM DETECTABLE EFFECT SIZE (MDES), AND SAMPLE SIZE CALCULATION FORMULA

General Notes

These notations apply to the below MDES formulas. *P* is the proportion of this sample that is treatment group. *n* is the average sample size for Level 1 (Students #). *J* is the average sample size for Level 2 (Classes #). *K* is the average sample size for Level 3 (School #). *L* is the total sample size for Level 4 (District #). ρ_2 (or ρ), ρ_3, and ρ_4 are unconditional intra-class correlation (ICC) at Levels, 2, 3, and 4, respectively. R_m^2 is the proportion of level-*m* variance explained by covariate at level *m* (*m* could be 1 – 4). R_{hT}^2 is the proportion of variance between level-*h* blocks on the treatment effect explained by block-level covariates (*h* could be 2 – 4). $\omega_h = \frac{\tau_{Th}^2}{\tau_h^2}$ indicates treatment effect heterogeneity (Hedges & Rhoads, 2010; Konstantopoulos, 2008b, 2009) across level-*h* block, which is proportion of the variance between level-*h* blocks on the treatment effect to the between level-*h*-block residual variance. For example, in two-level random effect block random assignment design (the model details are below), $\omega = \frac{\tau_{T2}^2}{\tau_2^2}$ indicates treatment effect heterogeneity across block. Note that $\rho\omega = \frac{\tau_{T2}^2}{\tau_2^2 + \sigma^2}$, which is effect size variability. The multiplier (M_v) for one-tailed test and two-tailed test are $t_\alpha + t_{1-\beta}$ and $t_{\alpha/2} + t_{1-\beta}$, respectively, with *v* degrees of freedom which is the function of the sample size and number of covariates depending on the study designs and analysis models.

1. Individual Random Assignment Design (IRA)

The treatment effect can be estimated by the ordinary least square model below:

$$Y_i = \beta_0 + \beta_1(TREATMENT)_i + \beta_2 X_i + e_i, \quad e_i \sim N(0, \sigma^2_{|X})$$

MDES formula is given by Bloom (2006), p.12:

$$MDES = M_{n-k^*-2}\sqrt{\frac{1-R_A^2}{nP(1-P)}}$$

The sample size ($n = n_T + n_C$) can be derived from the above formula as below:

$$n = \left(\frac{M_{n-k^*-2}}{MDES}\right)^2\left(\frac{1-R_A^2}{P(1-P)}\right)$$

P = the proportion of this sample that is randomized treatment, i.e., $n_T/(n_T + n_C)$. k^* = the number of covariates. $R_A^2 = 1-(\sigma^2_{|X}/\sigma^2)$, defined as the proportion of variance in the outcome predicted by covariates, X. σ^2 = variance in unconditional model (without any covariates). Multiplier for one-tailed test: $M_{n-k^*-2} = t_\alpha + t_{1-\beta}$ with n- k*-2 degrees of freedom. Multiplier for two-tailed test: $M_{n-k^*-2} = t_{\alpha/2} + t_{1-\beta}$ with n- k*-2 degrees of freedom. α is the type-I error, and β is the type-II error, i.e., (1-β) is the power.

Note that the multiplier, M_{n-k^*-2}, is a function of n, however, n can be solved through iterations.

2. Blocked Individual Random Assignment Design (BIRA)

Recall that in blocked individual random assignment design, treatment is at individual level (level 1).

Model 2.1. Two-level Blocked Individual Random Assignment Design, Constant Block Effect Model (BIRA2_1c).

The constant block effect model assumes that the treatment effect is constant across block. The statistical model only includes block dummy variables, which differentiate the intercepts.

The fixed block effect model assumes that each block has its own the treatment effect. The statistical model includes both block dummy variables and the interaction terms of block dummies and *TREATMENT* variable.

For the constant block effect model, within 2-level hierarchical linear model framework, we have:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TREATMENT)_{ij} + \beta_{2j}X_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2_{|X})$

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

Level 2: $\beta_{1j} = \gamma_{10}$

$$\beta_{2j} = \gamma_{20}$$

Reduced form: $Y_{ij} = \gamma_{00} + \gamma_{10}(TREATMENT)_{ij} + \gamma_{20}X_{ij} + \mu_{0j} + r_{ij}$

μ_{0j}, for $j \in \{1,2,...,J\}$, are associated with each block mean, constrained to have a mean of zero.

Bloom (2006, p.13) derived a MDES formula for the unconditional model (without covariate adjustment). The adapted MDES formula with covariate adjustment is:

$$MDES = M_{Jn-J-g_1^*-1}\sqrt{\frac{(1-R_1^2)}{JnP(1-P)}}$$

The level-2 sample size (*J*) can be derived from the above formula as below:

$$J = \left(\frac{M_{Jn-J-g_1^*-1}}{MDES}\right)^2\left(\frac{(1-R_1^2)}{nP(1-P)}\right)$$

Multiplier for one-tailed test: $M_{Jn-J-g^*-1} = t_\alpha + t_{1-\beta}$ with $Jn - J - g^* - 1$ degrees of freedom.

Multiplier for two-tailed test: $M_{Jn-J-g^*-1} = t_{\alpha/2} + t_{1-\beta}$ with $Jn - J - g^* - 1$ degrees of freedom.

R_1^2 is the proportion of pooled unexplained variation in the outcome predicted by the blocks and

covariates. *n* is the average number of individuals per block. *J* is the number of blocks. g_1^* is the number of covariates. *P* is the proportion of this sample that is treatment group (n_T / n).

Model 2.2. Two-level Blocked Individual Random Assignment Design, Fixed Block Effect Model (BIRA2_1f).

For the fixed block effect model, within 2-level hierarchical linear model framework, we have:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TREATMENT)_{ij} + \beta_{2j}X_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2_{|X})$

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

Level 2: $\beta_{1j} = \gamma_{10} + \mu_{1j}$

$$\beta_{2j} = \gamma_{20}$$

Reduced form: $Y_{ij} = \gamma_{00} + \gamma_{10}(TREATMENT)_{ij} + \gamma_{20}X_{ij} + \mu_{0j} + \mu_{1j}(TREATMENT)_{ij} + r_{ij}$

μ_{0j}, for $j \in \{1,2,...,J\}$, are fixed effects associated with each block mean, constrained to have a mean of zero; μ_{1j}, for $j \in \{1,2,...,J\}$, are fixed effects associated with each block treatment effect, constrained to have a mean of zero.

$$MDES = M_{Jn-2J-g_1^*}\sqrt{\frac{(1-R_1^2)}{JnP(1-P)}}$$

The level-2 sample size (*J*) can be derived from the above formula as below:

$$J = \left(\frac{M_{Jn-2J-g_1^*}}{MDES}\right)^2\left(\frac{(1-R_1^2)}{nP(1-P)}\right)$$

n = average number of individuals per block. g_1^* = number of level-1 covariates. R_1^2 = proportion of variance in the outcome predicted by blocks and level-1 covariates. *P* = the average proportion of this sample that is treatment group (n_T / n).

Model 2.3. Two-level Blocked Individual Random Assignment Design, Random Block Effect Model (BIRA2_1r).

Within 2-level hierarchical linear model framework, the unconditioanl model is:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TREATMENT)_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2)$

Level 2: $\begin{matrix} \beta_{0j} = \gamma_{00} + \mu_{0j} \\ \beta_{1j} = \gamma_{10} + \mu_{1j} \end{matrix} \qquad \begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_2^2 & \\ \tau_{2T2} & \tau_{T2}^2 \end{bmatrix} \right)$

Reduced form: $Y_{ij} = \gamma_{00} + \gamma_{10}(TREATMENT)_{ij} + \mu_{0j} + \mu_{1j}(TREATMENT)_{ij} + r_{ij}$.

The variance of *TREATMENT* derived by Raudenbush & Liu (2000) is as follows:

$$Var(\hat{\gamma}_{10}) = \frac{\tau_{T2}^2 + 4\sigma^2 / n}{J}.$$

$\rho = \dfrac{\tau_2^2}{\tau_2^2 + \sigma^2}$, unconditional intra-class coefficient (ICC).

$\omega = \dfrac{\tau_{T2}^2}{\tau_2^2}$ indicates treatment effect heterogeneity, which is the ratio of the variance of the treatment effect between blocks to the between-block residual variance. Note that $\rho\omega = \dfrac{\tau_{T2}^2}{\tau_2^2 + \sigma^2}$, which is effect size variability.

The conditioanl model is:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TREATMENT)_{ij} + \beta_{2j}X_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma_{|X}^2)$

Level 2: $\begin{matrix} \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + \mu_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + \mu_{1j} \\ \beta_{2j} = \gamma_{20} \end{matrix} \qquad \begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{2|W}^2 & \\ \tau_{2T2|W} & \tau_{T2|W}^2 \end{bmatrix} \right)$

$$MDES = M_{J-g^*-1}\sqrt{\frac{\rho\omega(1-R_{2T}^2)}{J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}}$$

The level-2 sample size (*J*) can be derived from the above formula as below:

$$J = \left(\frac{M_{J-g^*-1}}{MDES}\right)^2 \left(\rho\omega(1-R_{2T}^2) + \frac{(1-\rho)(1-R_1^2)}{P(1-P)n}\right)$$

The multiplier for a one-tailed test is: $M_{J-g^*-1} = t_\alpha + t_{1-\beta}$ with *J*- g^*-1 degrees of freedom. The multiplier for two-tailed test: $M_{J-g^*-1} = t_{\alpha/2} + t_{1-\beta}$ with *J*- g^*-1 degrees of freedom. *n* = average sample size for Level 1 (Students #). *P* = the average proportion of this sample that is treatment group (n_T / n). g^* = number of block-level covariates. $R_1^2 = 1 - \sigma_{|X}^2 / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R_{2T}^2 = 1 - \tau_{T2|W}^2 / \tau_{T2}^2$ indicates the proportion of variance between level-2 blocks on the treatment effect explained by level-2 covariates. When it is unclear how much the block-level covariate can reduce the block-treatment variance, it will be conservative to set $R_{2T}^2 = 0$.

Model 2.4. Three-level Blocked Individual Random Assignment Design, Random Block Effect Model (BIRA3_1r).

Within 3-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk}(TREATMENT)_{ijk} + \beta_{2jk}X_{ijk} + r_{ijk} \qquad r_{ijk} \sim N(0, \sigma_{|X}^2)$

Level 2:
$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \gamma_{01k}W_{jk} + \mu_{0jk} \\ \beta_{1jk} &= \gamma_{10k} + \gamma_{11k}W_{jk} + \mu_{1jk} \\ \beta_{2jk} &= \gamma_{20k} \end{aligned} \qquad \begin{pmatrix}\mu_{0jk}\\ \mu_{1jk}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}\tau_{2|W}^2 & \\ \tau_{2T2|W} & \tau_{T2|W}^2\end{bmatrix}\right)$$

Level 3:
$$\begin{aligned} \gamma_{00k} &= \xi_{000} + \xi_{001}V_k + \varsigma_{00k} \\ \gamma_{10k} &= \xi_{100} + \xi_{101}V_k + \varsigma_{10k} \\ \gamma_{01k} &= \xi_{010} \\ \gamma_{11k} &= \xi_{110} \\ \gamma_{20k} &= \xi_{200} \end{aligned} \qquad \begin{pmatrix}\varsigma_{00k}\\ \varsigma_{10k}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}\tau_{3|V}^2 & \\ \tau_{3T3|V} & \tau_{T3|V}^2\end{bmatrix}\right)$$

Based on the standard error of treatment effect estimate formula that derived by Hedges & Rhoads (2010) and Konstantopoulos (2008b), the MDES for 3-level blocked individual random assignment design with treatment at level 1 and random block effect model is as follows:

$$MDES = M_{K-g_3^*-1}\sqrt{\frac{\rho_3\omega_3(1-R_{3T}^2)}{K}+\frac{\rho_2\omega_2(1-R_{2T}^2)}{JK}+\frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)JKn}}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K=\left(\frac{M_{K-g_3^*-1}}{MDES}\right)^2\left(\rho_3\omega_3(1-R_{3T}^2)+\frac{\rho_2\omega_2(1-R_{2T}^2)}{J}+\frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)Jn}\right)$$

The multiplier for one-tailed test is: $M_{K-g_3^*-1}= t_\alpha + t_{1-\beta}$ with *K*- g_3^*-1 degrees of freedom. Multiplier for two-tailed test is: $M_{K-g_3^*-1}= t_{\alpha/2} + t_{1-\beta}$ with *K*- g_3^*-1 degrees of freedom. *J* = average sample size for Level 2 (Classes #). *n* = average sample size for Level 1 (Students #). *P* = the average proportion of this sample that is treatment group (n_T / n). $\rho_3 = \frac{\tau_3^2}{\tau_3^2+\tau_2^2+\sigma^2}$ is unconditional ICC at level 3. $\rho_2 = \frac{\tau_2^2}{\tau_3^2+\tau_2^2+\sigma^2}$ is unconditional ICC at level 2. τ_3^2 = level-3 variance (unconditional model). τ_2^2 = level-2 variance (unconditional model). σ^2 = individual-level variance (unconditional model). $\omega_3 = \frac{\tau_{T3}^2}{\tau_3^2}$ indicates treatment effect heterogeneity across level 3, which is the proportion of the variance between schools on the treatment effect to the between-school residual variance (unconditional model). $\omega_2 = \frac{\tau_{T2}^2}{\tau_2^2}$ indicates treatment effect heterogeneity across level 2, which is the proportion of the variance between classrooms on the treatment effect to the between-classroom residual variance (unconditional model). $\tau_{3|V}^2$ = level-3

variance conditional on level-3 covariate, *V*. $\tau^2_{2|W}$ = level-2 variance conditional on level-2 covariate, *W*. $\sigma^2_{|X}$ = individual-level variance conditional on level-1 covariate, *X*. $R^2_1 = 1 - \sigma^2_{|X} / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R^2_{2T} = 1 - \tau^2_{T2|W} / \tau^2_{T2}$ indicates the proportion of variance between level-2 blocks on the treatment effect explained by level-2 covariates. $R^2_{3T} = 1 - \tau^2_{T3|V} / \tau^2_{T3}$ indicates the proportion of variance between level-3 blocks on the treatment effect explained by level-3 covariates. g^*_3 = the number of group covariates used at level three. When it is unclear how much the block-level covariate can reduce the block-treatment variance, it will be conservative to set $R^2_{3T} = 0$; $R^2_{2T} = 0$.

Model 2.5. Four-level Blocked Individual Random Assignment Design, Random Block Effect Model (BCRA4_1r).

Within 4-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijkl} = \beta_{0jkl} + \beta_{1jkl}(TREATMENT)_{ijkl} + \beta_{2jkl} X_{ijkl} + r_{ijkl}, \quad r_{ijkl} \sim N(0, \sigma^2_{|X})$

Level 2:
$$\begin{aligned} \beta_{0jkl} &= \gamma_{00kl} + \gamma_{01kl} W_{jkl} + \mu_{0jkl} \\ \beta_{1jkl} &= \gamma_{10kl} + \gamma_{11kl} W_{jkl} + \mu_{1jkl} \\ \beta_{2jkl} &= \gamma_{20kl} \end{aligned}, \qquad \begin{pmatrix} \mu_{0jkl} \\ \mu_{1jkl} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau^2_{2|W} & \\ \tau_{2T2|W} & \tau^2_{T2|W} \end{bmatrix} \right)$$

Level 3:
$$\begin{aligned} \gamma_{00kl} &= \xi_{000l} + \xi_{001l} V_{kl} + \varsigma_{00kl} \\ \gamma_{10kl} &= \xi_{100l} + \xi_{101l} V_{kl} + \varsigma_{10kl} \\ \gamma_{01kl} &= \xi_{010l} \\ \gamma_{11kl} &= \xi_{110l} \\ \gamma_{20kl} &= \xi_{200l} \end{aligned}, \qquad \begin{pmatrix} \varsigma_{00kl} \\ \varsigma_{10kl} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau^2_{3|V} & \\ \tau_{3T3|V} & \tau^2_{T3|V} \end{bmatrix} \right)$$

$$\text{Level 4:}\quad \begin{aligned} \xi_{000l} &= \psi_{0000} + \psi_{0001} Z_l + \upsilon_{000l} \\ \xi_{001l} &= \psi_{0010} \\ \xi_{100l} &= \psi_{1000} + \psi_{1001} Z_l + \upsilon_{100l} \\ \xi_{101l} &= \psi_{1010} \\ \xi_{010l} &= \psi_{0100} \\ \xi_{110l} &= \psi_{1100} \\ \xi_{200l} &= \psi_{2000} \end{aligned}, \qquad \begin{pmatrix} \upsilon_{000l} \\ \upsilon_{100l} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau^2_{4|Z} & \\ \tau_{4T4|Z} & \tau^2_{T4|Z} \end{bmatrix} \right)$$

Following the same logic of 2- and 3- level block random assignment designs with treatment at level 1, the MDES formula for 4-level designs can be expressed as follows:

$$MDES = M_{L-g_4^*-1} \sqrt{\frac{\rho_4 \omega_4 (1-R^2_{4T})}{L} + \frac{\rho_3 \omega_3 (1-R^2_{3T})}{LK} + \frac{\rho_2 \omega_2 (1-R^2_{2T})}{LKJ} + \frac{(1-\rho_2-\rho_3-\rho_4)(1-R^2_1)}{P(1-P)LKJn}}$$

The level-4 sample size (L) can be derived from the above formula as below:

$$L = \left(\frac{M_{L-g_4^*-1}}{MDES} \right)^2 \left(\rho_4 \omega_4 (1-R^2_{4T}) + \frac{\rho_3 \omega_3 (1-R^2_{3T})}{K} + \frac{\rho_2 \omega_2 (1-R^2_{2T})}{KJ} + \frac{(1-\rho_2-\rho_3-\rho_4)(1-R^2_1)}{P(1-P)KJn} \right)$$

The multiplier for one-tailed test is: $M_{L-g_4^*-1} = t_\alpha + t_{1-\beta}$ with L- g_4^*-1 degrees of freedom. The multiplier for two-tailed test is: $M_{L-g_4^*-1} = t_{\alpha/2} + t_{1-\beta}$ with L- g_4^*-1 degrees of freedom. n = average number of individuals per level 2. P = average proportion of this sample that is treatment group (n_T / n). J = average sample size for Level 2 (Class #). K = average sample size for Level 3 (School #). $\rho_4 = \frac{\tau_4^2}{\tau_4^2 + \tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC at level 4. $\rho_3 = \frac{\tau_3^2}{\tau_4^2 + \tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC at level 3. $\rho_2 = \frac{\tau_2^2}{\tau_4^2 + \tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC at level 2. τ_4^2 = level-3 variance (unconditional model). τ_3^2 = level-3 variance (unconditional model). τ_2^2 = level-2 variance (unconditional model). σ^2 = individual-level

variance (unconditional model). $\omega_4 = \frac{\tau_{T4}^2}{\tau_4^2}$ indicates treatment effect heterogeneity across level 4, which is proportion of the variance between level-4 clusters on the treatment effect to the between level-4-cluster residual variance (unconditional model). $\omega_3 = \frac{\tau_{T3}^2}{\tau_3^2}$ indicates treatment effect heterogeneity across level 3, which is proportion of the variance between level-3 clusters on the treatment effect to the total between level-3-cluster residual variance (unconditional model). $\omega_2 = \frac{\tau_{T2}^2}{\tau_2^2}$ indicates treatment effect heterogeneity across level 2, which is proportion of the variance between level-2 clusters on the treatment effect to the total between level-2-cluster residual variance (unconditional model). $R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$, defined as the proportion of individual variance (at level one) predicted by covariates, *X*. $R_{2T}^2 = 1 - \tau_{T2|W}^2 / \tau_{T2}^2$ indicates the proportion of variance between level-2 blocks on the treatment effect explained by level-2 covariates. $R_{3T}^2 = 1 - \tau_{T3|V}^2 / \tau_{T3}^2$ indicates the proportion of variance between level-3 blocks on the treatment effect explained by level-3 covariates. $R_{4T}^2 = 1 - \tau_{T4|Z}^2 / \tau_{T4}^2$ indicates the proportion of variance between level-4 blocks on the treatment effect explained by level-4 covariates.

3. Simple Cluster Random Assignment Design (CRA)

Recall that in hierarchical random assignment designs, treatment is at top level.

Model 3.1. Two-level Cluster Random Assignment Design where treatment is at level 2 (CRA2_2r).

The treatment effect can be estimated by a 2-level hierarchical linear model:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma_{|X}^2)$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}W_j + \mu_{0j}, \mu_{0j} \sim N(0, \tau^2_{|W})$
$\beta_{1j} = \gamma_{10}$

Reduced form: $Y_{ij} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}W_j + \gamma_{10}X_{ij} + \mu_{0j} + r_{ij}$

The MDES formula from Bloom (2006, p.17) is:

$$MDES = M_{J-g^*-2}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}}$$

Sample attrition reduces statistical power by lowering the size of the analytic sample. For 2-level cluster random assignment design, attrition might occur at both levels. Suppose the retention rates (=1-attrition rates) at levels 1 and 2 are r_1 and r_2, respectively, the MDES formula containing the retention rates is:

$$MDES = M_{Jr_2-g^*-2}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)Jr_2} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jnr_2r_1}}$$

The level-2 sample size (*J*) can be derived from the above formula as below:

$$J = \left(\frac{M_{Jr_2-g^*-2}}{MDES}\right)^2\left(\frac{\rho(1-R_2^2)}{P(1-P)r_2} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)nr_2r_1}\right)$$

The multiplier for one-tailed test is: $M_{Jr_2-g^*-2} = t_\alpha + t_{1-\beta}$ with $Jr_2 - g^* - 2$ degrees of freedom. The multiplier for two-tailed test is: $M_{Jr_2-g^*-2} = t_{\alpha/2} + t_{1-\beta}$ with $Jr_2 - g^* - 2$ degrees of freedom.

τ^2 = Level-2 (between group-level) variance in unconditional model (without any covariates).

σ^2 = Level-1 (individual-level) variance in unconditional model (without any covariates). $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$, unconditional intra-class coefficient (ICC). $R_1^2 = 1 - (\sigma^2_{|X} / \sigma^2)$, defined as the proportion of individual variance at level one predicted by covariates, *X*. $R_2^2 = 1 - (\tau^2_{|W} / \tau^2)$,

defined as the proportion of group variance (at level two) predicted by covariates, *W*. *g** = the number of group covariates used. *P* = the proportion of this sample that is treatment group (J_T / J).

Model 3.2. Three-level Cluster Random Assignment Design where treatment is at level 3 (CRA3_3r).

The treatment effect can be estimated by a 3-level hierarchical linear model:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{ijk} + r_{ijk}, \quad r_{ijk} \sim N(0, \sigma^2_{|X})$

Level 2: $\begin{matrix} \beta_{0jk} = \gamma_{00k} + \gamma_{01k} W_{jk} + \mu_{0jk} \\ \beta_{1jk} = \gamma_{10k} \end{matrix}$, $\mu_{0jk} \sim N(0, \tau^2_{2|W})$

Level 3: $\begin{matrix} \gamma_{00k} = \xi_{000} + \xi_{001}(TREATMENT)_k + \xi_{002} V_k + \varsigma_{00k} \\ \gamma_{01k} = \xi_{010} \\ \gamma_{10k} = \xi_{100} \end{matrix}$, $\varsigma_{00k} \sim N(0, \tau^2_{3|V})$

Based on the variance (or standard error) of treatment effect estimate formula that derived by Hedges & Rhoads (2010), Konstantopoulos (2008a), Schochet (2008a), and Spybrook (2007), the MDES for 3-level cluster random assignment design is:

$$MDES = M_{K-g_3^*-2} \sqrt{\frac{\rho_3(1-R_3^2)}{P(1-P)K} + \frac{\rho_2(1-R_2^2)}{P(1-P)JK} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)JKn}}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K = \left(\frac{M_{K-g_3^*-2}}{MDES} \right)^2 \left(\frac{\rho_3(1-R_3^2)}{P(1-P)} + \frac{\rho_2(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)Jn} \right)$$

Multiplier for one-tailed test: $M_{K-g_3^*-2} = t_\alpha + t_{1-\beta}$ with *K*- g_3^*-2 degrees of freedom. Multiplier for two-tailed test: $M_{K-g_3^*-2} = t_{\alpha/2} + t_{1-\beta}$ with *K*- g_3^*-2 degrees of freedom. *J* = average sample

size for Level 2 (Classes #). $\rho_3 = \frac{\tau_3^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC at level 3. $\rho_2 = \frac{\tau_2^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC at level 2. τ_3^2 = level -3variance (unconditional model). τ_2^2 = level -2variance (unconditional model). σ^2 = individual-level variance (unconditional model). $R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$, defined as the proportion of individual variance (at level one) predicted by covariates, *X*. $R_2^2 = 1 - (\tau_{2|W}^2 / \tau_2^2)$, defined as the proportion of group variance (at level two) predicted by covariates, *W*. $R_3^2 = 1 - (\tau_{3|V}^2 / \tau_3^2)$, defined as the proportion of group variance (at level three) predicted by covariates, *V*. g_3^* = the number of group covariates used at level three. *P* = the proportion of this sample that is treatment group (K_T / K).

Model 3.3. Four-level Cluster Random Assignment Design where treatment is at level 4 (CRA4_4r).

The treatment effect can be estimated by a 4-level hierarchical linear model:

Level 1: $Y_{ijkl} = \beta_{0jkl} + \beta_{1jkl} X_{ijkl} + r_{ijkl}, \quad r_{ijkl} \sim N(0, \sigma_{|X}^2)$

Level 2:
$$\begin{aligned} \beta_{0jkl} &= \gamma_{00kl} + \gamma_{01kl} W_{jkl} + \mu_{0jkl} \\ \beta_{1jkl} &= \gamma_{10kl} \end{aligned}, \qquad \mu_{0jkl} \sim N(0, \tau_{2|W}^2)$$

Level 3:
$$\begin{aligned} \gamma_{00kl} &= \xi_{000l} + \xi_{001l} V_{kl} + \varsigma_{00kl} \\ \gamma_{01kl} &= \xi_{010l} \\ \gamma_{10kl} &= \xi_{100l} \end{aligned}, \qquad \varsigma_{00kl} \sim N(0, \tau_{3|V}^2)$$

Level 4:
$$\begin{aligned} \xi_{000l} &= \psi_{0000} + \psi_{0001} (TREATMENT)_l + \psi_{0002} Z_l + \upsilon_{000l} \\ \xi_{001l} &= \psi_{0010} \\ \xi_{010l} &= \psi_{0100} \\ \xi_{100l} &= \psi_{1000} \end{aligned}, \qquad \upsilon_{000l} \sim N(0, \tau_{4|Z}^2)$$

Following the same logic of 2- and 3- level cluster random assignment designs, the MDES formula for 4-level cluster random assignment designs can be expressed as follows:

$$MDES = M_{L-g_4^*-2}\sqrt{\frac{\rho_4(1-R_4^2)}{P(1-P)L}+\frac{\rho_3(1-R_3^2)}{P(1-P)KL}+\frac{\rho_2(1-R_2^2)}{P(1-P)JKL}+\frac{(1-\rho_2-\rho_3-\rho_4)(1-R_1^2)}{P(1-P)JKLn}}$$

The level-4 sample size (L) can be derived from the above formula as below:

$$L=\left(\frac{M_{L-g_4^*-2}}{MDES}\right)^2\left(\frac{\rho_4(1-R_4^2)}{P(1-P)}+\frac{\rho_3(1-R_3^2)}{P(1-P)K}+\frac{\rho_2(1-R_2^2)}{P(1-P)JK}+\frac{(1-\rho_2-\rho_3-\rho_4)(1-R_1^2)}{P(1-P)JKn}\right)$$

The multiplier for one-tailed test: $M_{L-g_4^*-2} = t_\alpha + t_{1-\beta}$ with L- g_4^*-2 degrees of freedom. The multiplier for two-tailed test: $M_{L-g_4^*-2} = t_{\alpha/2} + t_{1-\beta}$ with L- g_4^*-2 degrees of freedom. g_4^* = the number of Level 4 covariates. $\rho_4 = \frac{\tau_4^2}{\tau_4^2+\tau_3^2+\tau_2^2+\sigma^2}$, is the unconditional ICC at level 4. $R_4^2 = 1-(\tau_{4|Z}^2/\tau_4^2)$, defined as the proportion of level-4 variance predicted by covariates, Z. P = the proportion of this sample that is treatment group (L_T / L). All the other notations are same as in 3-level cluster random assignment design.

4. Blocked Cluster Random Assignment Design (BCRA)

In blocked cluster random assignment design, treatment is at sub-cluster level.

Model 4.1. Three-level Blocked Cluster Random Assignment Design (treatment at level 2), Fixed Block Effect Model (BCRA3_2f).

Within 3-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{ijk} + r_{ijk} \qquad r_{ijk} \sim N(0, \sigma_{|X}^2)$

Level 2: $\beta_{0jk} = \gamma_{00k} + \gamma_{01k}(TREATMENT)_{jk} + \gamma_{02k}W_{jk} + \mu_{0jk} \qquad \mu_{0jk} \sim N(0, \tau_{2|W}^2)$

$\beta_{1jk} = \gamma_{10k}$

Level 3:
$$\gamma_{00k} = \xi_{000} + \varsigma_{00k}$$
$$\gamma_{01k} = \xi_{010} + \varsigma_{01k}$$
$$\gamma_{02k} = \xi_{020}$$
$$\gamma_{10k} = \xi_{100}$$

ς_{00k}, for $k \in \{1,2,...,K\}$, are fixed effects associated with each block mean, constrained to have a mean of zero; ς_{01k}, for $k \in \{1,2,...,K\}$, are fixed effects associated with each block treatment effect, constrained to have a mean of zero.

Based on the variance of treatment effect estimate formula that derived by Spybrook (2007), the MDES for 3-level blocked cluster random assignment design with treatment at level 2 and fixed block effect model is:

$$MDES = M_{K(J-2)-g_2^*}\sqrt{\frac{\rho_2(1-R_2^2)}{P(1-P)JK} + \frac{(1-\rho_2)(1-R_1^2)}{P(1-P)JKn}}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K = \left(\frac{M_{K(J-2)-g_2^*}}{MDES}\right)^2\left(\frac{\rho_2(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho_2)(1-R_1^2)}{P(1-P)Jn}\right)$$

The multiplier for one-tailed test: $M_{K(J-2)-g_2^*} = t_\alpha + t_{1-\beta}$ with *K(J-2)-* g_2^* degrees of freedom. The multiplier for two-tailed test: $M_{K(J-2)-g_2^*} = t_{\alpha/2} + t_{1-\beta}$ with *K(J-2)-* g_2^* degrees of freedom. *J* = average sample size for Level 2 (Classes #). *P* = the proportion of this sample that is treatment group (J_T / *J*). $\rho_2 = \frac{\tau^2}{\tau^2 + \sigma^2}$, unconditional intra-class coefficient (ICC). τ^2 = Level-2 (between group-level) variance in unconditional model (without any covariates). σ^2 = Level-1 (individual-level) variance in unconditional model (without any covariates). $R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$, defined as the proportion of individual variance at level one predicted by covariates, *X*.

$R_2^2 = 1-(\tau_{|W}^2 / \tau^2)$, defined as the proportion of group variance (at level two) predicted by blocks and covariates, *W*. g_2^* = the number of Level 2 covariates.

Model 4.2. Three-level Blocked Cluster Random Assignment Design (treatment at level 2), Random Block Effect Model (BCRA3_2r).

Within 3-level hierarchical linear model framework:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{ijk} + r_{ijk} \qquad r_{ijk} \sim N(0, \sigma_{|X}^2)$

Level 2:
$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \gamma_{01k}(TREATMENT)_{jk} + \gamma_{02k} W_{jk} + \mu_{0jk} \qquad \mu_{0jk} \sim N(0, \tau_{2|W}^2) \\ \beta_{1jk} &= \gamma_{10k} \end{aligned}$$

Level 3:
$$\begin{aligned} \gamma_{00k} &= \xi_{000} + \xi_{001} V_k + \varsigma_{00k} \\ \gamma_{01k} &= \xi_{010} + \xi_{011} V_k + \varsigma_{01k} \qquad \begin{pmatrix} \varsigma_{00k} \\ \varsigma_{01k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{3|V}^2 & \\ \tau_{3T3|V} & \tau_{T3|V}^2 \end{bmatrix} \right) \\ \gamma_{02k} &= \xi_{020} \\ \gamma_{10k} &= \xi_{100} \end{aligned}$$

Based on the variance (or standard error) of treatment effect estimate formula that derived by Hedges & Rhoads (2010), Konstantopoulos (2008a), Schochet (2008a), and Spybrook (2007), the MDES for 3-level blocked cluster random assignment design with treatment at level-2 and random block effect model is:

$$MDES = M_{K-g_3^*-1} \sqrt{\frac{\rho_3 \omega (1-R_{3T}^2)}{K} + \frac{\rho_2 (1-R_2^2)}{P(1-P)JK} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)JKn}}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K = \left(\frac{M_{K-g_3^*-1}}{MDES} \right)^2 \left(\rho_3 \omega_3 (1-R_{3T}^2) + \frac{\rho_2 (1-R_2^2)}{P(1-P)J} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)Jn} \right)$$

The multiplier for one-tailed test is: $M_{K-g_3^*-1} = t_\alpha + t_{1-\beta}$ with *K*- g_3^*-1 degrees of freedom. The multiplier for two-tailed test is: $M_{K-g_3^*-1} = t_{\alpha/2} + t_{1-\beta}$ with *K*- g_3^*-1 degrees of freedom. *n* =

average number of individuals per level 2. J = average sample size for Level 2 (Classes #). P = the proportion of this sample that is treatment group (J_T / J). $\rho_3 = \frac{\tau_3^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$ is the correlation among students at the same school with different classes (unconditional model). $\rho_2 = \frac{\tau_2^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$ is the correlation among students at the same school with same classes (unconditional model). τ_3^2 = between group-level variance (level-3) (unconditional model). τ_2^2 = between group-level variance (level -2) (unconditional model). σ^2 = individual-level variance (unconditional model). $\omega = \frac{\tau_{T3}^2}{\tau_3^2}$ indicates treatment effect heterogeneity across block (school), which is the proportion of the variance between schools on the treatment effect to the between-school residual variance. $\tau_{3|V}^2$ = between group-level variance (level-3) (conditional model). $\tau_{2|W}^2$ = between group-level variance (level-2) (conditional model). $\sigma_{|X}^2$ = individual-level variance (conditional model). $R_1^2 = 1 - \sigma_{|X}^2 / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R_2^2 = 1 - \tau_{2|W}^2 / \tau_2^2$ indicates the proportion of variance between level-2 groups explained by level-2 covariates. $R_{3T}^2 = 1 - \tau_{T3|V}^2 / \tau_{T3}^2$ indicates the proportion of variance between level-3 blocks on the treatment effect explained by level-3 covariates. g_3^* = the number of group covariates used at level three.

Model 4.3. Four-level Blocked Cluster Random Assignment Design (treatment at level 2), Random Block Effect Model (BCRA4_2r).

Within 4-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijkl} = \beta_{0\,jkl} + \beta_{1\,jkl} X_{ijkl} + r_{ijkl}$, $r_{ijkl} \sim N(0, \sigma_{|X}^2)$

Level 2: $$\begin{aligned}\beta_{0jkl} &= \gamma_{00kl} + \gamma_{01kl}(TREATMENT)_{jkl} + \gamma_{02kl}W_{jkl} + \mu_{0jkl} \\ \beta_{1jkl} &= \gamma_{10kl}\end{aligned}, \qquad \mu_{0jkl} \sim N(0, \tau^2_{2|W})$$

Level 3: $$\begin{aligned}\gamma_{00kl} &= \xi_{000l} + \xi_{001l}V_{kl} + \varsigma_{00kl} \\ \gamma_{01kl} &= \xi_{010l} + \xi_{011l}V_{kl} + \varsigma_{01kl} \\ \gamma_{02kl} &= \xi_{020l} \\ \gamma_{10kl} &= \xi_{100l}\end{aligned}, \qquad \begin{pmatrix}\varsigma_{00kl} \\ \varsigma_{01kl}\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{bmatrix}\tau^2_{3|V} & \\ \tau_{3T3|V} & \tau^2_{T3|V}\end{bmatrix}\right)$$

Level 4: $$\begin{aligned}\xi_{000l} &= \psi_{0000} + \psi_{0001}Z_l + \upsilon_{000l} \\ \xi_{001l} &= \psi_{0010} \\ \xi_{010l} &= \psi_{0100} + \psi_{0101}Z_l + \upsilon_{010l} \\ \xi_{011l} &= \psi_{0110} \\ \xi_{020l} &= \psi_{0200} \\ \xi_{100l} &= \psi_{1000}\end{aligned}, \qquad \begin{pmatrix}\upsilon_{000l} \\ \upsilon_{010l}\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{bmatrix}\tau^2_{4|Z} & \\ \tau_{4T4|Z} & \tau^2_{T4|Z}\end{bmatrix}\right)$$

Following the same logic as in 3-level blocked individual random assignment design with treatment at level 1 and random block effect model:

$$MDES = M_{L-g_4^*-1}\sqrt{\frac{\rho_4\omega_4(1-R^2_{4T})}{L} + \frac{\rho_3\omega_3(1-R^2_{3T})}{LK} + \frac{\rho_2(1-R^2_2)}{P(1-P)LKJ} + \frac{(1-\rho_2-\rho_3-\rho_4)(1-R^2_1)}{P(1-P)LKJn}}$$

The level-4 sample size (*L*) can be derived from the above formula as below:

$$L = \left(\frac{M_{L-g_4^*-1}}{MDES}\right)^2\left(\rho_4\omega_4(1-R^2_{4T}) + \frac{\rho_3\omega_3(1-R^2_{3T})}{K} + \frac{\rho_2(1-R^2_2)}{P(1-P)KJ} + \frac{(1-\rho_2-\rho_3-\rho_4)(1-R^2_1)}{P(1-P)KJn}\right)$$

The multiplier for one-tailed test is: $M_{L-g_4^*-1} = t_\alpha + t_{1-\beta}$ with *L*- g_4^*-1 degrees of freedom. The multiplier for two-tailed test is: $M_{L-g_4^*-1} = t_{\alpha/2} + t_{1-\beta}$ with *L*- g_4^*-1 degrees of freedom. *n* = average number of individuals per level 2. *J* = average sample size for Level 2 (Class #). *P* = average proportion of this sample that is treatment group (J_T / J). *K* = average sample size for Level 3 (School #). *L* = total sample size for Level 4 (District #). ρ_2, ρ_3, and ρ_4 are the

unconditional ICCs at Levels 2, 3, and 4, respectively. $\omega_4 = \frac{\tau_{T4}^2}{\tau_4^2}$ indicates treatment effect heterogeneity across level 4, which is proportion of the variance between level-4 clusters on the treatment effect to the between level-4-cluster residual variance (unconditional model). $\omega_3 = \frac{\tau_{T3}^2}{\tau_3^2}$ indicates treatment effect heterogeneity across level 3, which is proportion of the variance between level-3 clusters on the treatment effect to the total between level-3-cluster residual variance (unconditional model). $R_1^2 = 1 - \sigma_{|X}^2 / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R_2^2 = 1 - \tau_{2|W}^2 / \tau_2^2$ indicates the proportion of variance between level-2 groups explained by level-2 covariates. $R_{3T}^2 = 1 - \tau_{T3|V}^2 / \tau_{T3}^2$ indicates the proportion of variance between level-3 blocks on the treatment effect explained by level-3 covariates. $R_{4T}^2 = 1 - \tau_{T4|Z}^2 / \tau_{T4}^2$ indicates the proportion of variance between level-4 blocks on the treatment effect explained by level-4 covariates. When it is unclear how much the block-level covariate can reduce the block-treatment variance, it will be conservative to set $R_{3T}^2 = 0$; $R_{4T}^2 = 0$.

Model 4.4. Four-level Blocked Cluster Random Assignment Design (treatment at level 3), Fixed Block Effect Model (BCRA4_3f).

Within 4-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijkl} = \beta_{0jkl} + \beta_{1jkl} X_{ijkl} + r_{ijkl}, \quad r_{ijkl} \sim N(0, \sigma_{|X}^2)$

Level 2: $\begin{aligned} \beta_{0jkl} &= \gamma_{00kl} + \gamma_{01kl} W_{jkl} + \mu_{0jkl} \\ \beta_{1jkl} &= \gamma_{10kl} \end{aligned}, \qquad \mu_{0jkl} \sim N(0, \tau_{2|W}^2)$

Level 3: $\begin{aligned} \gamma_{00kl} &= \xi_{000l} + \xi_{001l}(TREATMENT)_{kl} + \xi_{002l} V_{kl} + \varsigma_{00kl} \\ \gamma_{01kl} &= \xi_{010l} \\ \gamma_{10kl} &= \xi_{100l} \end{aligned}, \qquad \varsigma_{00kl} \sim N(0, \tau_{3|V}^2)$

$$\xi_{000l} = \psi_{0000} + \psi_{0001} Z_l + \upsilon_{000l}$$
$$\xi_{001l} = \psi_{0010} + \psi_{0011} Z_l + \upsilon_{001l}$$
Level 4: $\xi_{002l} = \psi_{0020}$
$$\xi_{010l} = \psi_{0100}$$
$$\xi_{100l} = \psi_{1000}$$

υ_{000l}, for $l \in \{1,2,...,L\}$, are fixed effects associated with each block mean, constrained to have a mean of zero; υ_{001l}, for $l \in \{1,2,...,L\}$, are fixed effects associated with each block treatment effect, constrained to have a mean of zero.

Using the same logic as in 3-level BRD with treatment at level 2 and fixed block effect model (Spybrook (2007), the MDES for 4-level blocked cluster random assignment design with treatment at level 3 and fixed block effect model is:

$$MDES = M_{L(K-2)-g_3^*} \sqrt{\frac{\rho_3(1-R_3^2)}{P(1-P)LK} + \frac{\rho_2(1-R_2^2)}{P(1-P)LKJ} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)LKJn}}$$

The level-4 sample size (*L*) can be derived from the above formula as below:

$$L = \left(\frac{M_{L(K-2)-g_3^*}}{MDES}\right)^2 \left(\frac{\rho_3(1-R_3^2)}{P(1-P)K} + \frac{\rho_2(1-R_2^2)}{P(1-P)KJ} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)KJn}\right)$$

The multiplier for one-tailed test is: $M_{L(K-2)-g_3^*} = t_\alpha + t_{1-\beta}$ with *L(K-2)-* g_3^* degrees of freedom. The multiplier for two-tailed test is: $M_{L(K-2)-g_3^*} = t_{\alpha/2} + t_{1-\beta}$ with *L(K-2)-* g_3^* degrees of freedom. *n* = average number of individuals per level 2. *P* = average proportion of this sample that is treatment group (n_T / n). *J* = average sample size for Level 2 (Class #). *K* = average sample size for Level 3 (School #). *P* = average proportion of this sample that is treatment group (K_T / K). g_3^* is the number of Level 3 covariates. $\rho_3 = \frac{\tau_3^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC

at level 3. $\rho_2 = \frac{\tau_2^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$ is the unconditional ICC at level 2. τ_3^2 = level-3 variance (unconditional model). τ_2^2 = level-2 variance (unconditional model). σ^2 = individual-level variance (unconditional model). $R_1^2 = 1 - \sigma_{|X}^2 / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R_2^2 = 1 - \tau_{2|W}^2 / \tau_2^2$ indicates the proportion of variance between level-2 groups explained by level-2 covariates. $R_3^2 = 1 - \tau_{3|V}^2 / \tau_3^2$ indicates the proportion of variance between level-3 groups explained by block and level-3 covariates.

Model 4.5. Four-level Blocked Cluster Random Assignment Design (treatment at level 3), Random Block Effect Model (BCRA4_3r).

Within 4-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijkl} = \beta_{0\,jkl} + \beta_{1\,jkl} X_{ijkl} + r_{ijkl}, \quad r_{ijkl} \sim N(0, \sigma_{|X}^2)$

Level 2: $\begin{aligned} \beta_{0\,jkl} &= \gamma_{00kl} + \gamma_{01kl} W_{jkl} + \mu_{0\,jkl} \\ \beta_{1\,jkl} &= \gamma_{10kl} \end{aligned}, \qquad \mu_{0\,jkl} \sim N(0, \tau_{2|W}^2)$

Level 3: $\begin{aligned} \gamma_{00kl} &= \xi_{000l} + \xi_{001l}(TREATMENT)_{kl} + \xi_{002l} V_{kl} + \varsigma_{00kl} \\ \gamma_{01kl} &= \xi_{010l} \\ \gamma_{10kl} &= \xi_{100l} \end{aligned}, \qquad \varsigma_{00kl} \sim N(0, \tau_{3|V}^2)$

Level 4: $\begin{aligned} \xi_{000l} &= \psi_{0000} + \psi_{0001} Z_l + \upsilon_{000l} \\ \xi_{001l} &= \psi_{0010} + \psi_{0011} Z_l + \upsilon_{001l} \\ \xi_{002l} &= \psi_{0020} \\ \xi_{010l} &= \psi_{0100} \\ \xi_{100l} &= \psi_{1000} \end{aligned}, \qquad \begin{pmatrix} \upsilon_{000l} \\ \upsilon_{001l} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{4|Z}^2 & \\ \tau_{4T\,4|Z} & \tau_{T\,4|Z}^2 \end{bmatrix} \right)$

Using the same logic as in 3-level BRD with treatment at level 2 and random block effect model, the MDES for 4-level blocked cluster random assignment design with treatment at level 3 and random block effect model is:

$$MDES = M_{L-g_4^*-1}\sqrt{\frac{\rho_4\omega_4(1-R_{4T}^2)}{L}+\frac{\rho_3(1-R_3^2)}{P(1-P)LK}+\frac{\rho_2(1-R_2^2)}{P(1-P)LKJ}+\frac{(1-\rho_2-\rho_3-\rho_4)(1-R_1^2)}{P(1-P)LKJn}}$$

The level-4 sample size (L) can be derived from the above formula as below:

$$L=\left(\frac{M_{L-g_4^*-1}}{MDES}\right)^2\left(\rho_4\omega_4(1-R_{4T}^2)+\frac{\rho_3(1-R_3^2)}{P(1-P)K}+\frac{\rho_2(1-R_2^2)}{P(1-P)KJ}+\frac{(1-\rho_2-\rho_3-\rho_4)(1-R_1^2)}{P(1-P)KJn}\right)$$

The multiplier for one-tailed test is: $M_{L-g_4^*-1} = t_\alpha + t_{1-\beta}$ with L- g_4^*-1 degrees of freedom. The multiplier for two-tailed test is: $M_{L-g_4^*-1} = t_{\alpha/2} + t_{1-\beta}$ with L- g_4^*-1 degrees of freedom. n = average number of individuals per level 2. J = average sample size for Level 2 (Class #). K = average sample size for Level 3 (School #). P = average proportion of this sample that is treatment group (K_T / K). ρ_2, ρ_3, and ρ_4 are unconditional ICCs at Levels 2, 3, and 4, respectively. $\omega_4 = \frac{\tau_{T4}^2}{\tau_4^2}$ indicates treatment effect heterogeneity across level 4, which is proportion of the variance between level-4 clusters on the treatment effect to the between level-4-cluster residual variance (unconditional model). $R_1^2 = 1 - \sigma_{|X}^2 / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R_2^2 = 1 - \tau_{2|W}^2 / \tau_2^2$ indicates the proportion of variance between level-2 groups explained by level-2 covariates. $R_3^2 = 1 - \tau_{3|V}^2 / \tau_3^2$ indicates the proportion of variance between level-3 groups explained by level-3 covariates. $R_{4T}^2 = 1 - \tau_{T4|Z}^2 / \tau_{T4}^2$ indicates the proportion of variance between level-4 blocks on the treatment effect explained by level-4 covariates. When it is unclear how much the block-level covariate can reduce the block-treatment variance, it will be conservative to set $R_{4T}^2 = 0$.

5. Regression Discontinuity Design (RD)

As discussed in the text, for the regression discontinuity design (RD), Schochet (2008b) summarized six types of commonly used cluster design based on the unit of treatment assignment and sampling framework. Based on the MDES formulas for the randomized experiments and design effect for their corresponding RD, the MDES for six types of RD can be shown below:

Model 5.1. Blocked individual regression discontinuity design with fixed effects (RD2_1f): analogous to the 2-level fixed effect blocked individual random assignment design (BIRA2_1f).

Within 2-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TREATMENT)_{ij} + \beta_{2j}X_{ij} + \beta_{3j}(Z_{ij} - Z_0) + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2_{|X})$

Level 2:
$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \mu_{1j} \\ \beta_{2j} &= \gamma_{20} \\ \beta_{3j} &= \gamma_{30} \end{aligned}$$

Z_{ij} is the assignment variable, and Z_0 is the cutoff score. μ_{0j}, for $j \in \{1,2,...,J\}$, are fixed effects associated with each block mean, constrained to have a mean of zero; μ_{1j}, for $j \in \{1,2,...,J\}$, are fixed effects associated with each block treatment effect, constrained to have a mean of zero.

$$MDES = M_{Jn-2J-g_1^*}\sqrt{\frac{D(1-R_1^2)}{JnP(1-P)}}$$

The level-2 sample size (*J*) can be derived from the above formula as below:

$$J = \left(\frac{M_{Jn-2J-g_1^*}}{MDES}\right)^2\left(\frac{D(1-R_1^2)}{nP(1-P)}\right)$$

n = average number of individuals per block. g_1^* = number of level-1 covariates. σ^2 = Level-1 (individual-level) variance (unconditional model). $R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$, indicating the proportion of level-1 variance explained by covariates. P = the average proportion of this sample that is treatment group (n_T / n). D is design effect (see Schochet, 2008b, Tables 4.1, 4.2, 4.3 for more information).

Note that the model could include quadratic or cubic terms of the assignment variable as well as the interaction terms of the assignment variable with treatment indicator. When higher order terms or/and interaction involve, the statistical power will decrease (Schochet, 2008b). This applies to all the below regression discontinuity designs.

Model 5.2. Blocked individual regression discontinuity design with random effects (RD2_1r): analogous to 2-level random effect blocked individual random assignment design with treatment at level 1 (BIRA2_1r).

Within 2-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}(TREATMENT)_{ij} + \beta_{2j}X_{ij} + \beta_{3j}(Z_{ij} - Z_0) + r_{ij}, \quad r_{ij} \sim N(0, \sigma_{|X}^2)$

Level 2:
$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + \mu_{1j} \\ \beta_{2j} &= \gamma_{20} \\ \beta_{3j} &= \gamma_{30} \end{aligned} \qquad \begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{2|W}^2 & \\ \tau_{2T2|W} & \tau_{T2|W}^2 \end{bmatrix} \right)$$

Z_{ij} is the assignment variable, and Z_0 is the cutoff score.

$$MDES = M_{J-g^*-1}\sqrt{\frac{\rho\omega(1-R_{2T}^2)}{J} + D\left(\frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}\right)}$$

The level-2 sample size (J) can be derived from the above formula as below:

$$J = \left(\frac{M_{J-g^*-1}}{MDES}\right)^2 \left(\rho\omega(1-R_{2T}^2) + D\left(\frac{(1-\rho)(1-R_1^2)}{P(1-P)n}\right)\right)$$

The multiplier for one-tailed test is: $M_{J-g^*-1} = t_\alpha + t_{1-\beta}$ with *J*- g^*-1 degrees of freedom. The multiplier for two-tailed test is: $M_{J-g^*-1} = t_{\alpha/2} + t_{1-\beta}$ with *J*- g^*-1 degrees of freedom. n = average sample size for Level 1 (Students #). *P* is the average proportion of this sample that is treatment group (n_T / n). g^* = number of block-level covariates. τ_{T2}^2 = variance of the treatment effect between blocks (unconditional model). τ_2^2 = Level-2 (between group-level) variance (unconditional model). σ^2 = Level-1 (individual-level) variance (unconditional model). ρ = $\frac{\tau_2^2}{\tau_2^2 + \sigma^2}$, unconditional intra-class coefficient (ICC). $\omega = \frac{\tau_{T2}^2}{\tau_2^2}$ indicates treatment effect heterogeneity, which is the ratio of the variance of the treatment effect between blocks to the between-block residual variance. $R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$, indicating the proportion of level-1 variance explained by covariates. $R_{2T}^2 = 1 - \tau_{T2|W}^2 / \tau_{T2}^2$ indicates the proportion of variance between level-2 blocks on the treatment effect explained by level-2 covariates. When it is unclear how much the block-level covariate can reduce the block-treatment variance, it will be conservative to set R_{2T}^2 = 0. *D* is design effect (see Schochet, 2008b, Tables 4.1, 4.2, 4.3 for more information). Note that the design effect only affects the level-1 term.

Model 5.3. Cluster regression discontinuity design sample with two levels of clustering and random effects (RDC_2r): analogous to 2-level simple cluster random assignment design with treatment at level 2 (model 3.1 CRA2_2r).

The treatment effect can be estimated by a 2-level hierarchical linear model:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2_{|X})$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(Z_j - Z_0) + \gamma_{03} W_j + \mu_{0j}, \mu_{0j} \sim N(0, \tau^2_{|W})$
$\beta_{1j} = \gamma_{10}$

Z_j is the assignment variable, and Z_0 is the cutoff score.

$$MDES = M_{J-g^*-2}\sqrt{D\left(\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}\right)}$$

The level-2 sample size (*J*) can be derived from the above formula as below:

$$J = \left(\frac{M_{J-g^*-2}}{MDES}\right)^2 D\left(\frac{\rho(1-R_2^2)}{P(1-P)} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)n}\right)$$

The multiplier for one-tailed test is: $M_{J-g^*-2} = t_\alpha + t_{1-\beta}$ with *J- g**-2 degrees of freedom. The multiplier for two-tailed test is: $M_{J-g^*-2} = t_{\alpha/2} + t_{1-\beta}$ with *J- g**-2 degrees of freedom. *J* = the total number of clusters. $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$, is the unconditional intra-class coefficient (ICC) . τ^2 = Level-2 (between group-level) variance (unconditional model). σ^2 = Level-1 (individual-level) variance (unconditional model). $R_1^2 = 1 - (\sigma^2_{|X} / \sigma^2)$, indicating the proportion of level-1 variance explained by covariates. $R_2^2 = 1 - (\tau^2_{|W} / \tau^2)$, indicating the proportion of level-2 variance explained by covariates. *g** = the number of group covariates used. *P* = the proportion of this sample that is treatment group (J_T / J). *D* = design effect (see Schochet, 2008b, Tables 4.1, 4.2, 4.3 for more information).

Model 5.4 (RDC_3r): analogous to 3-level simple cluster random assignment design with treatment at level 3 (Model 3.2 CRA3_3r).

The treatment effect can be estimated by a 3-level hierarchical linear model:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{ijk} + r_{ijk}, \quad r_{ijk} \sim N(0, \sigma^2_{|X})$

Level 2: $\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \gamma_{01k} W_{jk} + \mu_{0jk} \\ \beta_{1jk} &= \gamma_{10k} \end{aligned}, \qquad \mu_{0jk} \sim N(0, \tau^2_{2|W})$

Level 3: $\begin{aligned} \gamma_{00k} &= \xi_{000} + \xi_{001}(TREATMENT)_k + \xi_{002}(Z_k - Z_0) + \xi_{003} V_k + \varsigma_{00k} \\ \gamma_{01k} &= \xi_{010} \\ \gamma_{10k} &= \xi_{100} \end{aligned}, \ \varsigma_{00k} \sim N(0, \tau^2_{3|V})$

$$MDES = M_{K-g^*_3-2} \sqrt{D\left(\frac{\rho_3(1-R^2_3)}{P(1-P)K} + \frac{\rho_2(1-R^2_2)}{P(1-P)JK} + \frac{(1-\rho_2-\rho_3)(1-R^2_1)}{P(1-P)JKn}\right)}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K = \left(\frac{M_{K-g^*_3-2}}{MDES}\right)^2 D\left(\frac{\rho_3(1-R^2_3)}{P(1-P)} + \frac{\rho_2(1-R^2_2)}{P(1-P)J} + \frac{(1-\rho_2-\rho_3)(1-R^2_1)}{P(1-P)Jn}\right)$$

Z_k is the assignment variable, and Z_0 is the cutoff score. Multiplier for one-tailed test: $M_{K-g^*_3-2} = t_\alpha + t_{1-\beta}$ with *K*- g^*_3-2 degrees of freedom. Multiplier for two-tailed test: $M_{K-g^*_3-2} = t_{\alpha/2} + t_{1-\beta}$ with *K*- g^*_3-2 degrees of freedom. *J* = average sample size for Level 2 (Classes #). $\rho_3 = \frac{\tau^2_3}{\tau^2_3 + \tau^2_2 + \sigma^2}$ is the unconditional ICC at level 3. $\rho_2 = \frac{\tau^2_2}{\tau^2_3 + \tau^2_2 + \sigma^2}$ is the unconditional ICC at level 2. τ^2_3 = level -3variance (unconditional model). τ^2_2 = level -2variance (unconditional model). σ^2 = individual-level variance (unconditional model). $R^2_1 = 1 - (\sigma^2_{|X} / \sigma^2)$, defined as the proportion of individual variance (at level one) predicted by covariates, *X*. $R^2_2 = 1 - (\tau^2_{2|W} / \tau^2_2)$, defined as the proportion of group variance (at level two) predicted by covariates, *W*. $R^2_3 = 1 - (\tau^2_{3|V} / \tau^2_3)$, defined as the proportion of group variance (at level three) predicted by

covariates, *V*. g_3^* = the number of group covariates used at level three. *P* = the proportion of this sample that is treatment group (K_T / K). *D* = design effect (see Schochet, 2008b, Tables 4.1, 4.2, 4.3 for more information).

Model 5.5 (RD3_2f): analogous to 3-level fixed effect blocked cluster random assignment design with treatment at level 2 (Model 4.1 BCRA3_2f).

Within 3-level hierarchical linear model framework, the treatment effect can be estimated by:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{ijk} + r_{ijk} \qquad r_{ijk} \sim N(0, \sigma^2_{|X})$

Level 2:

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}(TREATMENT)_{jk} + \gamma_{02k}(Z_{jk} - Z_0) + \gamma_{03k} W_{jk} + \mu_{0jk}$$
$$\beta_{1jk} = \gamma_{10k}$$
$$\mu_{0jk} \sim N(0, \tau^2_{2|W})$$

$$\gamma_{00k} = \xi_{000} + \varsigma_{00k}$$
$$\gamma_{01k} = \xi_{010} + \varsigma_{01k}$$

Level 3: $\gamma_{02k} = \xi_{020}$
$\gamma_{03k} = \xi_{030}$
$\gamma_{10k} = \xi_{100}$

Z_{jk} is the assignment variable, and Z_0 is the cutoff score. ς_{00k} , for $k \in \{1,2,...,K\}$, are fixed effects associated with each block mean, constrained to have a mean of zero; ς_{01k} , for $k \in \{1,2,...,K\}$, are fixed effects associated with each block treatment effect, constrained to have a mean of zero.

$$MDES = M_{K(J-2)-g_2^*} \sqrt{D\left(\frac{\rho_2(1-R_2^2)}{P(1-P)JK} + \frac{(1-\rho_2)(1-R_1^2)}{P(1-P)JKn}\right)}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K = \left(\frac{M_{K(J-2)-g_2^*}}{MDES} \right)^2 D \left(\frac{\rho_2(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho_2)(1-R_1^2)}{P(1-P)Jn} \right)$$

The multiplier for one-tailed test is: $M_{K(J-2)-g_2^*} = t_\alpha + t_{1-\beta}$ with *K(J-2)-* g_2^* degrees of freedom. The multiplier for two-tailed test is: $M_{K(J-2)-g_2^*} = t_{\alpha/2} + t_{1-\beta}$ with *K(J-2)-* g_2^* degrees of freedom. *J* = average sample size for Level 2 (Classes #). *P* = the proportion of this sample that is treatment group (J_T / *J*). g_2^* = the number of Level 2 covariates. $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$, is the unconditional intra-class coefficient (ICC) . τ^2 = Level-2 (between group-level) variance (unconditional model). σ^2 = Level-1 (individual-level) variance (unconditional model). $R_1^2 = 1 - (\sigma_{|X}^2 / \sigma^2)$, indicating the proportion of level-1 variance explained by covariates. $R_2^2 = 1 - (\tau_{|W}^2 / \tau^2)$, indicating the proportion of level-2 variance explained by covariates. *D* = design effect (see Schochet, 2008b, Tables 4.1, 4.2, 4.3 for more information).

Model 5.6 (RD3_2r): analogous to 3-level random effect blocked cluster random assignment design with treatment at level 2 (model 4.2 BCRA3_2r).

Within 3-level hierarchical linear model framework:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{ijk} + r_{ijk} \qquad r_{ijk} \sim N(0, \sigma_{|X}^2)$

Level 2:
$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}(TREATMENT)_{jk} + \gamma_{02k}(Z_{jk} - Z_0) + \gamma_{03k} W_{jk} + \mu_{0jk}$$
$$\beta_{1jk} = \gamma_{10k}$$

$$\mu_{0jk} \sim N(0, \tau_{2|W}^2)$$

$$
\begin{aligned}
&\gamma_{00k} = \xi_{000} + \xi_{001}V_k + \varsigma_{00k} \\
&\gamma_{01k} = \xi_{010} + \xi_{011}V_k + \varsigma_{01k} \\
\text{Level 3: } &\gamma_{02k} = \xi_{020} \qquad \begin{pmatrix}\varsigma_{00k}\\ \varsigma_{01k}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}\tau^2_{3|V} & \\ \tau_{3T3|V} & \tau^2_{T3|V}\end{bmatrix}\right) \\
&\gamma_{03k} = \xi_{030} \\
&\gamma_{10k} = \xi_{100}
\end{aligned}
$$

Z_{jk} is the assignment variable, and Z_0 is the cutoff score.

$$MDES = M_{K-g_3^*-1}\sqrt{\frac{\rho_3\omega(1-R^2_{3T})}{K} + D\left(\frac{\rho_2(1-R^2_2)}{P(1-P)JK} + \frac{(1-\rho_2-\rho_3)(1-R^2_1)}{P(1-P)JKn}\right)}$$

The level-3 sample size (*K*) can be derived from the above formula as below:

$$K = \left(\frac{M_{K-g_3^*-1}}{MDES}\right)^2\left(\rho_3\omega_3(1-R^2_{3T}) + D\left(\frac{\rho_2(1-R^2_2)}{P(1-P)J} + \frac{(1-\rho_2-\rho_3)(1-R^2_1)}{P(1-P)Jn}\right)\right)$$

The multiplier for one-tailed test is: $M_{K-g_3^*-1} = t_\alpha + t_{1-\beta}$ with *K*- g_3^*-1 degrees of freedom. The multiplier for two-tailed test is: $M_{K-g_3^*-1} = t_{\alpha/2} + t_{1-\beta}$ with *K*- g_3^*-1 degrees of freedom. *n* = average number of individuals per level 2. *J* = average sample size for Level 2 (Classes #). *P* = the proportion of this sample that is treatment group (J_T / J). $\rho_3 = \frac{\tau_3^2}{\tau_3^2+\tau_2^2+\sigma^2}$ is the unconditional ICC at level 3. $\rho_2 = \frac{\tau_2^2}{\tau_3^2+\tau_2^2+\sigma^2}$ is the unconditional ICC at level 2. τ_3^2 = level -3variance (unconditional model). τ_2^2 = level -2variance (unconditional model). σ^2 = individual-level variance (unconditional model). $\omega = \frac{\tau^2_{T3}}{\tau_3^2}$ indicates treatment effect heterogeneity across block (school), which is the proportion of the variance between schools on the treatment effect to the between-school residual variance. $R_1^2 = 1-\sigma^2_{|X} / \sigma^2$ indicates the proportion of individual variance (at level one) predicted by covariates. $R_2^2 = 1-\tau^2_{2|W} / \tau_2^2$ indicates the proportion of

variance between level-2 groups explained by level-2 covariates. $R^2_{3T} = 1 - \tau^2_{T3|V} / \tau^2_{T3}$ indicates the proportion of variance between level-3 blocks on the treatment effect explained by level-3 covariates. g^*_3 = number of group covariates used at level three. *D* = design effect (see Schochet, 2008b, Tables 4.1, 4.2, 4.3 for more information). Note that design effect only affects Level-1 and Level-2 terms.

6. Interrupted Time-Series Design:

The time-series design (Bloom, 2003; Quint, Bloom, Black, & Stephens, 2005) compares student scores before and after a school-wide intervention while modeling the underlying pre-intervention trend over time and the departures from that trend during the post-intervention years. Those departures from the pre-intervention trend provide the estimates of the intervention effects.

The analytic model for this time-series is multilevel with students nested within cohorts within each school. Each school provides estimates of intervention effects; schools thus constitute blocks in this design. Following Bloom (2003)'s suggestion, the analysis can proceed as follows:

Series 1: These time series analyses assess whether there are improvements in the scores on each of the performance variables for any series of grade level cohorts in the intervention schools after the intervention begins. Using student-level scores on the respective performance variable in each cohort as the dependent variable, this analysis examines differences between the years prior to the intervention and those afterwards. Two cohort-level (level 2) variables model change over time in the analysis. One variable, denoted in the models below as *T*, indicates where in the time series each student cohort was located. The other variable, denoted in the models below as D_t, is a set of dummy variables that indicates that the cohort was in the intervention (D_t=1) or not (D_t=0) for each year after the onset of the intervention.

First, we present a simple model to estimate the program effect at the second implementation year with 4 waves of baseline information by assuming that the intervention effect is constant across schools:

Level 1 (student)

$$(1)\quad Y_{ijk} = \beta_{0jk} + \beta_{1jk}C_{ijk} + r_{ijk}\ \ r_{ijk} \sim N(0, \sigma^2)$$

Level 2 (cohort: random effect)

$$(2)\quad \begin{aligned} \beta_{0jk} &= \beta_{00k} + \beta_{01k}T_{jk} + \beta_{02k}X_{jk} + \sum_{t=0}^{2}\beta_{0(3+t)k}(D_t)_{jk} + \mu_{0jk} \quad \begin{pmatrix}\mu_{0jk}\\ \mu_{1jk}\end{pmatrix} \sim N\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}\tau_{00} & \\ \tau_{01} & \tau_{11}\end{pmatrix} \\ \beta_{1jk} &= \beta_{10k} + \mu_{1jk} \end{aligned}$$

Level 3 (school: constant effect)

$$(3)\quad \begin{aligned} \beta_{00k} &= \gamma_{000} + \sum_m \gamma_{00m}S_m \\ \beta_{01k} &= \gamma_{010} \\ \beta_{02k} &= \gamma_{020} \\ \beta_{0(3+t)k} &= \gamma_{0(3+t)0}, t = 0,1,2. \\ \beta_{10k} &= \gamma_{100} \end{aligned}$$

Reduced Form:

$$(4)\quad \begin{aligned} Y_{ijk} &= \gamma_{000} + \sum_m \gamma_{00m}S_m + \gamma_{010}T_{jk} + \gamma_{020}X_{jk} + \sum_{t=0}^{2}\gamma_{0(3+t)0}(D_t)_{jk} + \gamma_{100}C_{ijk} \\ &+ \mu_{0jk} + C_{ijk}\mu_{1jk} + r_{ijk} \end{aligned}$$

Y_{ijk} = Score for student, *i* at time, *j* in school, *k*. C_{ijk} = Covariate for student, *i* at time, *j* in school, *k*. S_m = Dummy variable indicating School *m* (representing the blocking factor) .

X_{jk} = Cohort-level covariate. T_{jk} = Test year for student *i* (ranging from - 4 through + 2). D_t = Dummy variables indicating the intervention status for cohort at follow-up year *t* = 0, 1, and 2. γ_{000} = Grand mean score for students at the baseline in the reference school. γ_{00m} = Difference in the grand mean score for students at baseline in the other schools comparing with the

reference school. γ_{010} = Linear trend. γ_{020} = Slopes of the mean cohort achievement. $\gamma_{0(3+t)0}$ = Deviations from trend for follow-up year t = 0, 1, and 2. γ_{100} = Coefficient of the student-level covariate. μ_{0jk} = Random error term for cohort at time, j in school, k. μ_{1jk} = Random error term in the slope of student-level covariate for cohort at time, j in school, k. r_{ijk} = Random error term for student, i at time, j in school, k.

The model above will estimate the program effects ($\gamma_{0(3+t)0}$) in terms of deviations from the trend for the follow-up years, and it assumes that the school effect is constant. This is a strong assumption; and the schools could have differential effects. Bloom (2003) proposed a school fixed effects model by estimating the program effects separately by school, then average them by weighting them equally. The model above can be extended to capture Bloom's idea by adding school dummies in the level-3 equation (3) above to predict level-2 parameters, i.e., adding the interaction terms of school dummies and the other variables in the reduced model. The reduced form model can be expressed as:

$$
\begin{aligned}
Y_{ijk} &= \gamma_{000} + \sum_m \gamma_{00m} S_m + \gamma_{010} T_{jk} + \gamma_{020} X_{jk} + \sum_{t=0}^{2} \gamma_{0(3+t)0} (D_t)_{jk} + \gamma_{100} C_{ijk} \\
&+ \sum_m \gamma_{01m} S_m * T_{jk} + \sum_m \gamma_{02m} S_m * X_{jk} + \sum_m \sum_{t=0}^{2} \gamma_{0(3+t)m} S_m * (D_t)_{jk} + \sum_m \gamma_{10m} S_m * C_{ijk} \\
&+ \mu_{0jk} + C_{ijk} \mu_{1jk} + r_{ijk}
\end{aligned} \qquad (5)
$$

γ_{01m} = Difference in linear trends for the other school comparing with the reference school. γ_{02m} = Difference in slopes of the mean cohort scores for the other schools comparing with the reference school. $\gamma_{0(3+t)m}$ = Difference in deviations from trends for follow-up year t = 0, 1, and 2 for the other schools comparing with the reference school. γ_{10m} = Difference in coefficient of the

student-level covariate for the other schools comparing with the reference school. The other notations are same as in Models 1 – 4.

The estimates of the program effects in the follow-up years in terms of deviations from the baseline year are $\gamma_{0(3+t)0}$, and ($\gamma_{0(3+t)0} + \gamma_{0(3+t)m}$) for the reference schools and all other schools. The simple average of these estimates will be the estimate of the intervention effect within the framework of the school fixed effect model.

Series 2: An important limitation of the above analyses is that observed differences in the student performance scores before and after the intervention begins in a school do not necessarily indicate that the differences were due to the intervention. This issue can be addressed by including similar schools that do not receive the intervention as comparison schools, i.e. schools located in the same districts and thereby subject to the same local context as the intervention schools. This comparison will serve to account for differences in a school's pattern of achievement that might be attributed to factors other than the intervention. Expanded from the models in series 1, an additional dummy variable will indicate whether the student's cohort was located in an intervention school or comparison school. The interaction of intervention and school type will examine whether treatment cohorts in program schools outperformed cohorts in comparison schools in the same time period.

Similar to the analysis in Series 1, we will start from the simple model—assuming the school effect is constant. By adding a dummy variable indicating if a school is a program school or a comparison school in equations (3) and (4), the reduced model is:

$$
\begin{aligned}
(6)\quad & Y_{ijk} = \gamma_{000} + \sum_{m} \gamma_{00m} S_m + \gamma_{010} T_{jk} + \gamma_{020} X_{jk} + \sum_{t=0}^{2} \gamma_{0(3+t)0} (D_t)_{jk} + \gamma_{100} C_{ijk} \\
& + \gamma_{00p} (INT)_k + \sum_{t=0}^{2} \gamma_{0(3+t)1} (INT)_k * (D_t)_{jk} + \mu_{0jk} + C_{ijk} \mu_{1jk} + r_{ijk}
\end{aligned}
$$

$(INT)_k$ = Dummy variables indicating if school, *k*, is an intervention school or comparison school. $\gamma_{0(3+t)0}$ = Deviations from trend for follow-up year *t* = 0, 1, and 2 for the comparison schools. The other notations are same as in Models 1 – 4.

The term $\gamma_{0(3+t)1}$ represents the average intervention effect in terms of the difference in deviations from the trend for the follow-up years between the intervention schools and comparison schools.

A more complicated fixed school effects model that permits the schools to have differential effects extends equation (6) by adding the interaction terms of school dummies and the other variables. The reduced model is:

$$
\begin{aligned}
Y_{ijk} &= \gamma_{000} + \sum_m \gamma_{00m} S_m + \gamma_{010} T_{jk} + \gamma_{020} X_{jk} + \sum_{t=0}^{2} \gamma_{0(3+t)0} (D_t)_{jk} + \gamma_{100} C_{ijk} \\
&+ \sum_m \gamma_{01m} S_m * T_{jk} + \sum_m \gamma_{02m} S_m * X_{jk} + \sum_m \sum_{t=0}^{2} \gamma_{0(3+t)m} S_m * (D_t)_{jk} + \sum_m \gamma_{10m} S_m * C_{ijk} \\
&+ \gamma_{00p} (INT)_k + \sum_{t=0}^{2} \gamma_{0(3+t)1} (INT)_k * (D_t)_{jk} + \mu_{0jk} + C_{ijk} \mu_{1jk} + r_{ijk}
\end{aligned}
\tag{7}
$$

The notations in equation (7) are same as in equations (5) and (6). $\gamma_{0(3+t)1}$ are the parameter estimates of interest representing the average intervention effect in terms of the difference in deviations from the trend for the follow-up years between the program schools and comparison schools.

Using a cohort random effects model (with school as a constant effect), Bloom (1999, 2003) presented a formula to calculate the Minimum Detectable Effect Size (MDES). An adapted MDES formula including comparison schools and covariates is below:

$$
MDES = \frac{M}{\sqrt{m}} \sqrt{1 + \frac{1}{p}} \sqrt{\frac{1}{n} + \frac{\rho(1 - R_2^2)}{1 - \rho}} \sqrt{1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}}
$$

The number of program schools which get treatment (*m*) can be derived from the above formula as below:

$$m = \left(\frac{M}{MDES}\right)^2 \left(1 + \frac{1}{p}\right)\left(\frac{1}{n} + \frac{\rho(1-R_2^2)}{1-\rho}\right)\left(1 + \frac{1}{T} + \frac{(t_f - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}\right)$$

where *n* is the number of students per school, *T* is the number of baseline year, t_f is the follow-up year of interest, $\bar{t}$ is the mean baseline year, ρ is the conditional intra-class correlation for cohorts (proportion of total variance of between years), *M* is multiplier, for one-tailed test: $M_{m*T-g^*-1} = t_\alpha + t_{1-\beta}$ with $m*T-g^*-1$ degrees of freedom. For two-tailed test: $M_{m*T-g^*-1} = t_{\alpha/2} + t_{1-\beta}$ with $m*T-g^*-1$ degrees of freedom, and *p* is the ratio of the number of comparison schools to the number of program schools.