Title:

Design and Analysis of Cluster Randomized Trials

Authors and Affiliations:

Wei Li*
University of Florida
PO Box 117049
Gainesville, FL 32611
Phone: 352-273-4332
Email: wei.li@coe.ufl.edu
ORCID: http://orcid.org/0000-0002-4082-9118

Yanli Xie
University of Cincinnati
Cincinnati, Ohio 45221
xieyl@ucmail.uc.edu

Dung Pham
Western Michigan University
3571 Sangren Hall
Kalamazoo, Michigan 49008
dungthuy.pham@wmich.edu

Nianbo Dong*
University of North Carolina at Chapel Hill
116 Peabody Hall, CB 3500
Chapel Hill, NC 27599
Phone: (919)843-9553
dong.nianbo@gmail.com
ORCID: http://orcid.org/0000-0003-0128-7106

Jessaca Spybrook
Western Michigan University
3571 Sangren Hall
Kalamazoo, Michigan 49008
Phone: (269) 387-3889
jessaca.spybrook@wmich.edu
ORCID: http://orcid.org/0000-0003-1768-6187

Benjamin Kelcey
University of Cincinnati
3311B RECCENTER
Cincinnati, Ohio 45221
Tel: 513-556-3608

ben.kelcey@gmail.com

*Corresponding Authors
Correspondence concerning this article should be addressed to Nianbo Dong, Peabody Hall CB 3500, Chapel Hill, NC 27599-3500, Email: nianbo.dong@unc.edu and Wei Li, College of Education, University of Florida, PO Box 117049, Email: wei.li@coe.ufl.edu.

Abstract:

Cluster randomized trials (CRTs) are commonly used to evaluate the causal effects of educational interventions, where the entire clusters (e.g., schools) are randomly assigned to treatment or control conditions. This study introduces statistical methods for designing and analyzing two-level (e.g., students nested within schools) and three-level (e.g., students nested within classrooms nested within schools) CRTs. Specifically, we utilize hierarchical linear models (HLMs) to account for the dependency of the intervention participants within the same clusters, estimating the average treatment effects (ATEs) of educational interventions and other effects of interest (e.g., moderator and mediator effects). We demonstrate methods and tools for sample size planning and statistical power analysis. Additionally, we discuss common challenges and potential solutions in the design and analysis phases, including the effects of omitting one level of clustering, non-compliance, threats to external validity, and cost-effectiveness of the intervention. We conclude with some practical suggestions for CRT design and analysis, along with recommendations for further readings.

Keywords:

Cluster randomized trials, Hierarchical linear models, Statistical power analysis, Causal inference

Statements and Declarations:

This project is supported by the National Science Foundation [[#1552535, #1760884, #1913563, and #2000705]. The opinions expressed herein are those of the authors and not the funding agency.

Design and Analysis of Cluster Randomized Trials

Randomized controlled trials have been widely used to assess the efficacy of educational interventions worldwide. For example, a recent systematic review (Connolly et al., 2018) identified 1,017 unique randomized trials in education for the period 1980–2016, with over half of all the randomized trials identified being conducted in the U.S. The common use of randomized trials in education in the U.S., especially in the past two decades, could be attributed to several factors related to the creation of the Institute of Education Sciences (IES). In 2002, Congress passed the Education Sciences Reform Act, establishing IES as a scientific and funding agency for high-quality education research. The IES, through its initiative known as What Works Clearinghouse (WWC), considers randomized trials the most rigorous research design to test the efficacy of educational interventions. The National Center for Education Research (NCER) and National Center for Special Education Research (NCSER) provide funding for randomized trials in education primarily in the form of grants. IES also offers professional development for graduate students and early-career researchers through its pre and postdoctoral training programs in universities and for established researchers through its summer training institutes (see Hedges & Schauer (2018) for details) to increase the capacity of the field to conduct randomized trials. Other government agencies such as the National Science Foundation (NSF) and the Office of Investment and Innovation (OII) and foundations such as the William T. Grant Foundation and the Spencer Foundation have supported the use of randomized trials in program evaluation. These determined efforts have led to a surge in the number of randomized trials of educational interventions in the U.S., which has provided the field with abundant and rigorous information about "what works" to inform education policy, practice, and research.

Educational interventions usually involve nested data structure (e.g., students nested within classrooms nested within schools), and the treatment can be at any level. Specifically, when the treatment is at the lower or middle level (e.g., classroom or student level), it is usually called a multisite randomized controlled trial, while when the treatment is at the top level (e.g., school level), it is commonly called a cluster randomized trial (CRT). This paper focuses on CRTs frequently used in education, where the whole clusters (e.g., schools) rather than individuals (e.g., students) are randomly assigned to a treatment or control group. For example, IES has funded more than 250 experimental studies, most of which are CRTs (Spybrook et al., 2020). In Asia, CRTs have also been used to assess the effectiveness of educational programs. For example, Mo et al. (2014) evaluated the effect of a computer-assisted learning (CAL) tutoring program on the mathematics achievement of third- and fifth-graders in China, where they randomly assigned a total of 72 elementary schools to the tutoring program or the business-as-usual condition.

CRTs in education often include two levels of clustering (e.g., students nested within schools) or three levels of clustering (e.g., students nested within teachers and teachers nested within schools). Students' outcomes (e.g., test scores) within the same clusters are dependent due to shared experiences or resources (Raudenbush, 1997; Raudenbush & Bryk, 2002). Educational researchers face many challenges in addressing data dependencies during the stages of designing and analyzing CRTs. For example, traditional statistical methods (e.g., OLS regression, ANOVA, etc.) cannot be used to estimate the treatment effects in CRTs because these methods assume that each observation is independent of the others, an assumption that is violated in CRTs. The hierarchical linear models (HLMs) consider the dependence among students by allowing a random effect for each cluster (Raudenbush & Bryk, 2002) and thus are typically used

in education. Research methodologists in education have utilized HLMs to develop statistical methods for planning and analyzing CRTs, such as statistical power analysis (e.g., Dong & Maynard, 2013; Raudenbush et al., 2011; Tipton & Miller, 2016). They also implemented power and sample size computation methods into user-friendly tools and made them freely available online through the support of federal agencies and foundations. However, these methods and tools are not widely applied outside the U.S. For example, the inclusion of a power analysis is uncommon for CRTs conducted in Asia. A meta-analysis of randomized experiments of educational interventions aimed to improve elementary school student outcomes in developing countries, including East Asian and Pacific and South Asian countries (McEwan, 2015), found that just fewer than half of experiments reported conducting an a priori power analysis to determine the sample size necessary for the study to detect an effect of a given magnitude at a specific power level.

This paper aims to introduce the recent development of statistical methods and tools for the design and analysis of CRTs in education. We focus on two- and three-level designs with continuous outcomes and begin by introducing the statistical models, assumptions, and sample size planning methods using a working example. Then, we discuss the methods and tools for addressing some common challenges educational researchers meet during the design and analysis phases. We conclude with some practical suggestions regarding CRT design and analysis and recommendations for further readings.

Methods

A Working Example

This primer will introduce and illustrate the CRT design and analysis methods through a working example as follows. Suppose a research team wants to evaluate whether a computer-

assisted learning (CAL) tutoring program could improve fifth graders' mathematics achievement through an experimental design. They have the option to utilize either a completely randomized design or a cluster randomized design. As shown in Figure 1, in a completely randomized design, students are individually assigned to either a treatment or control group. In contrast, a cluster randomized design assigns entire schools to either the treatment or control group, with all students within a school receiving the same condition. The research team may opt for a CRT due to its advantages over a completely randomized trial (Donner et al., 2000; Hayes & Moulton, 2017). For instance, CRTs can reduce the risk of contamination between treatment and control groups, which can occur when individuals in the same school influence each other. Furthermore, it might be more practical or feasible to implement the CAL tutoring program to entire schools rather than to random students within schools. During the design phase, they need to determine (1) whether to conduct a two-or three-level design, (2) the effect of interest to estimate, (3) the statistical models used to estimate these effects, and (4) the sample size at each level to guarantee adequate power to detect the effect of interest. This section will discuss the statistical methods and tools to answer these questions.

Potential outcome framework

The causal effects of educational interventions are commonly defined through the potential outcome (PO) framework or Rubin Causal Model (RCM; Imbens & Rubin, 2015). We use a three-level design (e.g., students nested classrooms within schools) as an example to illustrate how RCM defines the causal effects, and note that similar definitions apply to two-level designs.

Specifically, the causal effect of the CAL tutoring program for each student – the individual treatment effect (ITE) is defined as

$$ITE = Y_{ijk}(1) - Y_{ijk}(0), \quad (1)$$

where $Y_{ijk}(1)$ represents the potential outcome for student i in classroom j in school k if assigned to the treatment group and $Y_{ijk}(0)$ represents the potential outcome if assigned to the control group. The fundamental problem of causal inference is that we cannot observe both potential outcomes because each student can only be assigned to one group. Therefore, one of the two potential outcomes must be missing in practice, and ITE cannot be estimated. Instead, the RCM seeks to estimate the average treatment effect (ATE) for a particular population of participants.

For CRTs, the ATEs can be defined at the student, classroom, or school level. For example, the school-level ATE is defined as

$$ATE = E[Y_k(1) - Y_k(0)], \quad (2)$$

where $Y_k(1)$ and $Y_k(0)$ represent school-level potential outcomes for treatment and control groups, which are the school-level averages of student-level potential outcomes that are taken over the distribution of schools, and $E[Y_k(1) - Y_k(0)]$ represents the expected value of the difference between two potential outcomes. Similarly, the classroom-level ATE is defined as $E[Y_{jk}(1) - Y_{jk}(0)]$, where $Y_{jk}(1)$ and $Y_{jk}(0)$ represent classroom-level potential outcomes, which are the classroom-level averages of student-level potential outcomes that are taken over the distribution of classrooms, and the student-level ATE is defined as $E[Y_{ijk}(1) - Y_{ijk}(0)]$.

It requires two key assumptions to estimate the ATEs at different levels: the unconfounded assignment and the Stable Unit Treatment Value Assumption (SUTVA; Imbens & Rubin, 2015). The unconfounded assignment assumption is guaranteed because of the random assignment. The SUTVA indicates no interference at different levels in identifying the ATEs at corresponding levels. It might be violated at a particular level but is reasonable at other levels

(Rhoads & Li, 2023). For example, CRTs administer the intervention at the cluster level, thereby requiring SUTVA to hold at this level, but not necessarily at the individual level. When treatments have significant peer effects within the same cluster, CRTs enable researchers to identify a combined effect that includes both the direct impact of treatment on individuals and the peer effects, assuming SUTVA is maintained at the cluster level. Similarly, to identify the student-level ATE, the potential outcome for student *i* in classroom *j* from school *k* should not depend on what treatment other students receive, which might not hold because of the peer effects (or spillover effects) within the same classroom and school. Recent literature has developed methods of identifying student-level ATE with spillover effects (e.g., DiTraglia et al., 2023; Vazquez-Bare, 2023). It also should be noted that, when the unconfounded assignment and SUTVA hold, the ATEs at different levels are equal if the number of students per classroom and the number of classrooms per school are the same (Rhoads & Li, 2023). For the designs with the sample sizes varying among classes and schools, we must assume constant treatment effects across students and classrooms (i.e., no treatment heterogeneity across students and classrooms) to make the ATEs at different levels equal.

Statistical Model

HLMs are widely accepted as a common approach to estimating ATEs for CRTs in education (Raudenbush & Bryk, 2002). Specifically, for a two-level design (e.g., students nested with schools), assume there are *J* level-2 units (e.g., schools) in the study, and each level-2 unit has *n* level-1 units (e.g., students). The proportion of level-2 units assigned to the treatment condition is *P*. Then, the research team can use a two-model with a level-2 random intercept to estimate the ATE of the CAL tutoring program.

Students Level:

$$Y_{ij} = \beta_{0j} + \mathbf{B_{1j}X_{ij}} + e_{ij}, e_{ij} \sim N(0, \sigma^2_{|\mathbf{X}}), \quad (3)$$

School Level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \mathbf{\Gamma_{02}W_j} + u_{0j}, u_{0j} \sim N\left(0, \tau^2_{|T,\mathbf{W}}\right), \quad (4)$$

$$\mathbf{B_{1j}} = \mathbf{\Gamma_{10}}, \quad (5)$$

where Y_{ij} is the mathematics scores for student i in school j, $\mathbf{X_{ij}}$ is a column vector of baseline student-level covariates (e.g., gender, race, prescore, etc.), $\mathbf{B_{1j}}$ is a row vector of student-level coefficients that are assumed to be fixed among schools T_j is a binary treatment indicator variable coded as one for the students assigned to the treatment group and zero otherwise, $\mathbf{W_j}$ is a column vector of baseline level-2 covariates (e.g., enrollment, location, type, etc.), $\mathbf{\Gamma_{02}}$ is a row vector of school-level coefficients, and $\mathbf{\Gamma_{10}}$ is a vector of school-level fixed effects. Note that the baseline covariates (e.g., $\mathbf{X_{ij}}$ and $\mathbf{W_j}$) should be correlated with the outcome (Y_{ij}) to improve the accuracy of the estimated treatment effects. We usually assume that the error term at level 1 (e_{ij}) follows a normal distribution with a mean zero and a constant conditional variance $\sigma^2_{|\mathbf{X}}$ and the random effect at level 2 (u_{0j}) follows a normal distribution with a mean zero and a constant conditional variance $\tau^2_{|T,\mathbf{W}}$. To get an unbiased estimate of the treatment effect (or a causal effect), the binary treatment indicator should be independent of the level-1 error (e_{ij}) and level-2 random effect (u_{0j}), which is met by design because of random assignment (Raudenbush & Bryk, 2002). Therefore, the parameter of interest—γ_{01}—represents the ATE of the CAL tutoring program.

Similarly, for a three-level design, assume that there are *K* level-3 units (e.g., schools), each level-3 unit has *J* level-2 units (e.g., classrooms), and each level-2 unit has *n* level-1 units (e.g., students). The proportion of level-3 units assigned to the treatment condition is *P*. The

research team can utilize a three-level model with level-2 and level-3 random intercepts to estimate the ATE of the CAL tutoring program for three-level designs, namely

Students Level:

$$Y_{ijk} = \beta_{0jk} + \mathbf{B_{1jk}X_{ijk}} + e_{ijk},\ e_{ijk} \sim N\left(0, \sigma^2_{|\mathbf{X}}\right), \tag{6}$$

Class Level

$$\beta_{0jk} = \pi_{00k} + \mathbf{\Pi_{01k}W_{jk}} + r_{0jk},\ r_{0jk} \sim N\left(0, \tau^2_{|\mathbf{W}}\right), \tag{7}$$

$$\mathbf{B_{1jk}} = \mathbf{\Pi_{10k}}, \tag{8}$$

School Level:

$$\pi_{00k} = \gamma_{000} + \gamma_{001}\, T_k + \mathbf{\Gamma_{002}Z_k} + u_{00k},\ u_{00k} \sim N\left(0, \omega^2_{|T,\mathbf{Z}}\right), \tag{9}$$

$$\mathbf{\Pi_{01k}} = \mathbf{\Gamma_{010}}, \tag{10}$$

$$\mathbf{\Pi_{10k}} = \mathbf{\Gamma_{100}}, \tag{11}$$

where Y_{ijk} is the mathematics score for student i in classroom j in school k, $\mathbf{X_{ijk}}$ is a column vector of baseline student-level covariates, $\mathbf{B_{1jk}}$ is a row vector of student-level coefficients that are assumed to be fixed at the class and school levels, $\mathbf{W_{jk}}$ is a column vector of baseline level-2 covariates (e.g., teacher's gender, age, education level, class size, etc.), $\mathbf{\Pi_{01k}}$ is a row vector of class-level coefficients that are assumed to be fixed at the school level, T_k is a binary treatment indicator variable, $\mathbf{Z_k}$ is a column vector of baseline school-level covariates (e.g., enrollment, location, type, etc.), and $\mathbf{\Gamma_{002}}$ is a row vector of school-level coefficients. $\mathbf{\Gamma_{010}}$ and $\mathbf{\Gamma_{100}}$ are vectors of school-level fixed effects. The error term at level 1 (e_{ijk}) and the random effects at levels 2 and 3 (r_{0jk} and u_{0k}) are assumed to follow normal distributions with mean zeros and constant conditional variances. Because of random assignment, the treatment indicator (T_k) is independent of level-1 error (e_{ij}) and the random effects at levels 2 and 3 (r_{0jk} and u_{0k}). Thus,

the fixed effect associated with the treatment indicator —γ_{001}—represents the ATE of the CAL tutoring program.

Hypothesis Testing, Effect Size, and Power Analysis

Educational researchers are commonly interested in two general questions when evaluating educational interventions: whether the intervention is effective and how large the ATE is. To answer the first question, the hypothesis testing approach is utilized. Specifically, we can use a *t*-test to examine the null hypothesis $H_0: \gamma = 0$, where $\gamma = \gamma_{01}$ or γ_{001} for two- or three-level designs, respectively, and the alternative hypothesis is $H_a: \gamma \neq 0$. The *t*-statistic is the ratio of the estimate of the treatment effect ($\hat{\gamma}_{01}$ or $\hat{\gamma}_{001}$) to its standard error (SE). Educational researchers can conduct HLM analyses using major statistical software (e.g., R, SAS, SPSS, Stata, etc.). In addition, educational researchers usually transform the treatment effects to effect sizes to interpret their magnitude and compare them among alternative programs. Effect size is a standardization of the treatment effect and is defined as

$$\delta = \frac{\gamma}{\sqrt{Var(Y)}}, \tag{12}$$

where $Var(Y)$ is the total variance of the outcome. Therefore, the effect size can be interpreted in terms of the standard deviation of the outcome. Table 1 provides the exact formulas of effect size for two- and three-level CRTs.

When a non-significant treatment effect is observed, there are three potential explanations: (1) the intervention was indeed ineffective, (2) the intervention was effective, but the study was not adequately powered to detect the effect, and (3) the study had sufficient power to detect the true effect, but due to the specific outcome of the randomization (e.g., "unhappy" randomization), the estimated effect turned out to be quite small (biased to 0) and thus

insignificant. The probability of detecting the effect of interest when the alternative hypothesis is true is known as statistical power.

In the design phase, educational researchers need to figure out how many students, classrooms, and schools are needed to ensure adequate power (e.g., ≥ 0.8) to detect the treatment effect if it truly exists. One way to determine the sample size of a CRT is to calculate power. As we discussed above, under the normal distribution assumptions, we can test the null hypothesis $\gamma = 0$ using a *t*-test. Assuming the alternative hypothesis is true, the test statistic follows a non-central *t*-distribution with a non-centrality parameter:

$$\hat{\lambda} = \frac{\hat{\gamma}}{\sqrt{Var(\hat{\gamma})}}. \tag{13}$$

Under these specifications, the statistical power of a two-tailed test is (note $t_0 = t_{1-\frac{\alpha}{2},df}$)

$$Power = 1 - \mathrm{H}\,[t_0,\ df,\ \hat{\lambda}] + \mathrm{H}\,[-t_0,\ df,\ \hat{\lambda}], \tag{14}$$

where *df* is the degrees of freedom for the test and $\mathrm{H}(t_0,\ df,\ \hat{\lambda})$ is the cumulative distribution function of the non-central *t*-distribution with a non-centrality parameter $\hat{\lambda}$.

Another approach to deciding the sample size is to compute the minimum detectable effect size (MDES), which is defined as the smallest effect that has an acceptable chance (e.g., power > 0.8) of producing a treatment estimate that is statistically significant at the significance criterion (Bloom, 1995). In general, the MDES can be computed using the following formula:

$$MDES = M_v * \sqrt{\frac{Var(\gamma)}{Var(Y)}}, \tag{15}$$

where M_v represents the sum of two *t* quantiles (Bloom, 1995). In particular, for two-tailed tests, which are usually applied, $M_v = t_{\alpha/2} + t_{1-\beta}$, where α represents the Type I error and β represents the Type II error for the tests (Bloom, 1995; Dong & Maynard, 2013).

Table 2 summarizes the computation formulas of the non-centrality parameters, MDES, and degrees of freedom for the analysis of main effects in two- and three-level designs based on prior literature (e.g., Bloom, 1995; Dong & Maynard, 2013; Konstantopoulos, 2008; Raudenbush, 1997). In general, power and MDES are determined by sample size at each level and a set of design parameters. Table 1 provides the meanings and computation formulas of these parameters, including the nested effects of the outcome and the covariates effect. Specifically, the nested effects are represented by the intra-class correlation coefficients (ICCs), indicating the proportion of variance at the second or third level to the total variance of the outcome, and it is negatively correlated with power. The covariate effects are represented by the proportion of variance explained by the covariates at a particular level. Please note that we include both unconditional models that do not include any covariates at any level and models with covariates (i.e., conditional models) in Table 2 to facilitate defining effect size and design parameters. In practice, it is recommended to incorporate covariates into the HLMs because many prior studies have shown that they can increase statistical power considerably (e.g., Bloom, Richburg-Hayes, & Black, 2007; Dong & Maynard, 2013; Dong et al., 2016; Konstantopoulos, 2008; Raudenbush, Martinez, & Spybrook, 2007). In addition, the sample sizes at the cluster level impact power and MDES more than the sample sizes at lower levels (e.g., Konstantopoulos, 2008), and thus, education researchers may consider sampling more schools than classes or students if the study budget and resources permit.

Demonstration of Power and MDES Computation

Numerous software options are available to help applied researchers conduct power analyses for CRTs, such as Optimal Design (Spybrook et al., 2011). We utilize PowerUp! tools (e.g., PowerUp!, PowerUpR, PyPowerUp!, etc.) because they are easy to use and freely available

from the Causal Evaluation website (https://www.causalevaluation.org/). In particular, the tool – PowerUp! (Dong & Maynard, 2013) is primarily designed to facilitate the MDES and the sample size computations for a given research design and analysis. It is implemented through the user-friendly Microsoft Excel program and includes multiple worksheets, each specific to a particular design and analysis. To use this tool, users can follow a four-step procedure: (1) choose a study design (e.g., cluster random assignment design); (2) specify the number of clustering levels within the study (e.g., two or three); (3) specify the calculator to compute the MDES or the sample size for a given MDES; (4) specify the values of design parameters and statistical significance tests – the yellow highlighted parameters in the worksheet. Once users input these parameters, the PowerUp! automatically computes MDES or sample size. For example, suppose the research team would like to know the MDES for a two-level CRT that investigates the CAL tutoring program. In that case, they need first to choose "Simple Cluster Random Assignment," select two levels and MDES, and then input all the required parameters highlighted in grey. In general, the research team can rely on three strategies to identify the reference values of the design parameters: consulting prior literature for similar studies, using large databases to estimate these parameters, and calculating these parameters from a pilot study (Spybrook et al., 2016). For example, prior studies have documented empirical values of the design parameters for student achievement (Hedberg & Hedges, 2014; Hedges & Hedberg, 2007). Table 3 provides an example of MDES computation that randomly assigns 40 schools (i.e., level-2 units; $J = 40$) into treatment or control conditions in equal proportion ($P = 0.5$) with 100 students from each school ($n = 100$). In this example, the research team specified an alpha level of 0.05, a two-tailed test, 0.8 power, an ICC of 0.23, 50% variance explained at both levels ($R_1^2 = R_2^2 = 0.5$), and one covariate at level 2. After inputting these parameters highlighted in yellow, PowerUp! returned an estimate of the MDES equal to 0.314 (shown in Bold at the bottom of the worksheet).

If instead, the research team would like to compute the statistical power for a particular design and set of design parameters, they can use the PowerUpR package and its Shiny App (Bulus et al., 2022). For applied researchers not familiar with R programming, the PowerUpR Shiny App is recommended. Like using PowerUp!, users must first select a particular design (e.g., a three-level CRT) and then input the reference values of design parameters for power computation. Figure 2 shows an example output of PowerUpR Shiny App for a three-level design, where the research team randomly assigns 50 schools ($K = 50$) into treatment and control conditions in equal proportion ($P = 0.5$) with four classes from each school ($j = 4$) and 25 students from each class ($n = 25$). In this example, the research team specified an alpha level of 0.05, a two-tailed test, an effect size of 0.25 ($\delta = 0.25$), an ICC of 0.15 at level 3 ($\rho_3 = 0.15$), an ICC of 0.05 at level 2 ($\rho_2 = 0.05$), R^2 of 0.5 at all levels ($R_1^2 = R_2^2 = R_3^2 = 0.5$), and one covariate at level 3. PowerUpR Shiny provided the power estimate (statistical power = 0.843), degrees of freedom of the test, and Type I and II error rates.

Moderation and Mediation Analyses

Besides the ATE of an intervention, educational researchers and policymakers are often interested in exploring for whom, under what conditions, and through what mechanisms an intervention works or fails. For example, when evaluating the CAL tutoring program, besides the ATE (i.e., main effect), the research team might want to explore whether the treatment effect varies among subgroups of students, classrooms (or teachers), and schools with different characteristics (e.g., students' gender, teachers' educational level, school locale, etc.). These characteristics used to define subgroups are called moderators and could potentially alter the treatment effects (Baron & Kenny, 1986; Raudenbush & Liu, 2000). The moderator effects can be evaluated through HLMs with an interaction term between the treatment and moderator

variables. For example, to assess whether or not the effect of the CAL tutoring program varies between public and private schools in their two-level CRT, the research team can add a treatment by school type interaction term into the second-level model, namely

Students Level:

$$Y_{ij} = \beta_{0j} + \mathbf{B_{1j}X_{ij}} + e_{ij},\ e_{ij} \sim N\left(0, \sigma^2_{|\mathbf{X}}\right), \tag{16}$$

School Level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\, T_j + \mathbf{\Gamma_{02}W_j} + \gamma_{03} S_j + \gamma_{04} T_j S_j + u_{0j},\ u_{0j} \sim N\left(0, \tau_{|T,\mathbf{W},S}\right), \tag{17}$$

$$\mathbf{B_{1j}} = \mathbf{\Gamma_{10}}, \tag{18}$$

where S_j is a binary private school indicator coded as one if for private schools and zero for public schools. The parameter of interest is γ_{04}, indicating the differences in treatment effects between private and public schools. We recommend centering the private school indicator (S_j) so that γ_{01} still represents the ATE; otherwise, it represents the treatment effect for private schools (i.e., $S_j = 0$). Also, the research team can add a cross-level interaction to the two-level HLMs to evaluate whether the treatment effect varies between male and female students, namely

Students Level:

$$Y_{ij} = \beta_{0j} + \beta_{1j} F_{ij} + \mathbf{B_2X_{ij}} + e_{ij},\ e_{ij} \sim N\left(0, \sigma^2_{|F,\mathbf{X}}\right), \tag{19}$$

School Level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\, T_j + \mathbf{\Gamma_{02}W_j} + u_{0j},\ u_{0j} \sim N\left(0, \tau_{|T,\mathbf{W}}\right), \tag{20}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\, T_j + u_{1j},\ u_{1j} \sim N\left(0, \tau_{|T}\right), \tag{21}$$

$$\mathbf{B_{2j}} = \mathbf{\Gamma_{20}}, \tag{22}$$

where F_{ij} is a binary female indicator coded as one for females and zero otherwise. The parameter of interest now is γ_{11}, indicating the difference in the treatment effects between male

and female students. Note that Equation 19 assumes that the gender gap at level 1 (i.e., β_{1j}) varies among schools after accounting for treatment effects, represented by the random effects u_{1j}. The research team can also assume that the gender gap does not vary among schools and exclude the random effects from Equation 19. Similarly, the research team can evaluate whether the treatment effect varies among classrooms or teachers under a three-level design.

Mediation analysis has been widely used to investigate the intermediate mechanism through which a treatment effect is transmitted (Baron & Kenny, 1986). The intermediate variable is called a mediator. Mediation analysis can inform not only how a treatment works but also identify gaps in a theory (Hayes, 2013; MacKinnon, Fairchild, & Fritz, 2007; Kelcey et al., 2020). For example, if some prior theory indicates a direct relationship between self-esteem and student achievement but the mediation analysis identifies a significant indirect effect through a mediator (e.g., students' motivation), it suggests that motivation plays a vital role in explaining how self-esteem influences student achievement. Thus, researchers can conclude that the theory could be further advanced to include the role of motivation. Mediation analysis is commonly applied through path analyses and has several modeling options and strategies depending on the level of mediators (e.g., Kelcey et al., 2017). For example, the research team can utilize mediation analysis to examine whether the CAL remedial tutoring program improves students' learning through increasing student motivation, where the treatment is at the school level (level 2), and the outcome (e.g., test scores) and mediator (i.e., student motivation) are measured at the student level (level 1). Such analysis can be implemented via a 2-1-1 mediation model, where "2" represents the level of treatment, the first "1" represents the level of mediator, and the second "1" represents the level of outcome. Figure 3 illustrates the paths of this mediation model. Specifically, path coefficient *a* represents the effects of the treatment (CAL tutoring) on the

mediator (a measure of students' motivation), path coefficient *b* represents the effect of the mediator on students' mathematics achievement, path coefficient *c'* represents the direct effect of the treatment on students' mathematics achievement, and path *c* represents the total effect of the treatment on the outcome. In simple mediation analyses, the total or main effect can be decomposed as

Total effect = indirect effect + direct effect or *c* = *ab* + *c'*.

The research team must choose appropriate statistical models to address the potential clustering effect because of the nested data structure when estimating these effects. For this 2-1-1 model, they can estimate the total effect of the CAL program on students' mathematics achievement using a two-level HLM that is comparable in structure to Equations 3, 4, and 5. It should be noted that the mediator should be excluded from the level-1 model when estimating the total effect. The research team examines the effects of the treatment on student motivation (path *a* in Figure 3) using a two-level model for the mediator, namely

Students Level:

$$m_{ij} = \beta_{0j} + \mathbf{B_{1j}X_{ij}} + \zeta_{ij},\ \zeta_{ij} \sim N\left(0, \eta^2_{|\mathbf{X}}\right), \tag{23}$$

School Level:

$$\beta_{0j} = \pi_{00} + aT_j + \mathbf{\Gamma_{02}W_j} + u_{0j},\ u_{0j} \sim N\left(0, \tau_{|T,\mathbf{W}}\right), \tag{24}$$

$$\mathbf{B_{1j}} = \mathbf{\Gamma_{10}}, \tag{25}$$

where m_{ij} represents a measure of motivation for student *i* in school *j*, ζ_{ij} is an error term that follows a normal distribution with a mean zero and a constant conditional variance $\eta^2_{|\mathbf{X}}$, and *a* is the effect of the CAL tutoring program on the mediator. All the other terms have been defined previously. To estimate the conditional relationship between the mediator and outcome, path *b* in Figure 3, the research team needs to use another two-level model, namely

Students Level:

$$Y_{ij} = \beta_{0j} + \beta_{1j} m_{ij} + \mathbf{B_{2j} X_{ij}} + e_{ij},\ e_{ij} \sim N\left(0, \sigma^2_{|\mathbf{X},m}\right), \tag{26}$$

School Level:

$$\beta_{0j} = \alpha_{00} + b\bar{m}_j + c'T_j + \mathbf{A_{02} Z_j} + u_{0j},\ u_{0j} \sim N\left(0, \tau^2_{0\,|T,\mathbf{Z}}\right), \tag{27}$$

$$\beta_{1j} = \alpha_{10}, \tag{28}$$

$$\mathbf{B_{2j}} = \mathbf{A_{20}}, \tag{29}$$

where β_{1j} is student-level coefficient that is assumed to be fixed at level 2, c' is the direct effect, $\bar{m}_j$ is the average student motivation across students in school j with coefficient b as the total conditional relationship between the mediator and mathematics achievement. The mediation effect (indirect effect) could be calculated as $a \times b$. Similar path analyses can be applied for level-2 mediators or three-level designs (see Kelcey, Spybrook, & Dong, 2019; Kelcey et al., 2017; Kelcey et al., 2021; Kelcey et al., 2021).

Under this formulation, we assume that exposure to the treatment does not moderate the relationship between the mediator and outcome. However, this assumption can be relaxed by adding an interaction between the treatment and mediator in the outcome model (e.g., Kelcey et al., 2016). Additionally, this approach takes up four primary assumptions. First, the identifiability of the mediation effect requires nonzero and overlapping probabilities for exposure to a treatment. Second, identifiability requires the SUTVA assumption. As discussed above, SUTVA includes that, for example, one person's potential response (on the mediator or outcome) does not depend on another person's treatment (or mediator) status. In multilevel settings, this is complicated by the clustering or shared experience of students within schools. Prior literature has developed several approaches to relax this assumption by tracking or approximating the nature of that influence (e.g., responses depend on the cluster mean values of

the mediator; Hong, 2005). In our example above, the influence of a student's peers on the outcome may also operate through the average motivation of students within a school—that is, the collective peer motivation level within a school creates a contextual factor that also shapes the potential outcomes (e.g., Kelcey et al., 2017). Put differently, a student's potential outcome also depends on the motivation of schoolmates in addition to treatment exposure and individual motivation.

The third primary assumption to identify mediation effects under this approach is sequential ignorability. This assumption precludes unmeasured confounding of the treatment-outcome relationship, treatment-mediator relationship, and the mediator-outcome relationship (Vanderweele, 2015). The first and second components of this assumption are typically fulfilled through the random assignment. However, because mediator values are not randomly assigned, ensuring the mediator-outcome relationship is unconfounded requires adjustment for covariates that potentially confound the relationship. In our running example, we need to consider students' efficacy and learning engagement as well as other variables that are likely to influence motivation and the outcome. In addition, identifiability of the mediation effect typically draws on the assumption that there are no variables downstream of the treatment that are influenced by the treatment and subsequently confound the mediator-outcome relationship. Prior literature has developed several approaches to relax this assumption under different conditions and through alternative estimands (e.g., Imai & Yamamoto, 2013; Vansteelandt, 2017). More generally, literature has developed an array of sensitivity analyses to probe the robustness of results to violations of each of these assumptions (e.g., Vanderweele, 2015)

In terms of planning studies to detect effects, the research team needs to conduct power analyses during the design phase to make sure the sample size is large enough to detect

moderator or mediator effects. More specifically, prior research has widely underscored the importance of supplementing main effect power analyses with moderation- and mediation-specific power analyses because they typically propose different sample sizes and resource allocations (e.g., Cox & Kelcey, 2019; IES RFA, 2023; Kelcey et al., 2017; Sim et al., 2022). For example, professional standards across an array of fields require power analyses for moderation and mediation effects regardless of whether these components are primary, secondary, or exploratory in nature (e.g., Fairchild et al., 2017; IES RFA, 2023; Vo et al., 2019). Careful consideration of all effects in the planning stages helps ensure rigorous designs that address the full breadth of evidence sought. Even if researchers eventually privilege the power to detect one effect (e.g., main) over another effect (e.g., mediation) in the final design, these choices will be guided by informed tradeoffs.

Prior studies (e.g., Dong et al., 2018, 2021; Kelcey et al., 2021; Spybrook et al., 2016) have developed the methods of computing statistical power and MDES for mediation and moderation analysis and have implemented them into user-friendly tools – PowerUp!-Moderator (Dong et al., 2017a) and PowerUp!-Mediator (Dong et al., 2017b) for applied researchers to conduct power analysis, which are also freely available the Causal Evaluation website (https://www.causalevaluation.org/).

Common Challenges

In this section, we discuss the potential solutions to some common challenges facing educational researchers when designing CRTs, including the effects of omitting one level of clustering, non-compliance, heterogeneous variance, blocking, threats to external validity, and cost-effectiveness of the intervention.

Omitting Middle-Level Information. It is recommended that the research design and analysis should be consistent with the nested data structure. For example, CRTs in education commonly involve students, classrooms, and schools, where the treatment is at the school level and there are more than one classroom sampled from each school. Therefore, a three-level design is recommended, and a three-level HLM should be used to estimate the ATE. However, the middle level (e.g., classroom) might be missing in practice because the available datasets might only identify the schools that students attend but not the classrooms in which they are taught (Zhu et al., 2012). In that case, the research team has to use a two-level HLM instead. Prior studies (e.g., Moerbeek, 2004) have shown that, for models that only include the treatment indicator without any covariates at any level (i.e., an unconditional model), if the classroom level is ignored, part of classroom-level variance will shift up to the school level, and the rest will shift down to the student level. As a result, the classroom-level variance is accounted for (e.g., Moerbeek, 2004), and the main treatment effect and its SE will be the same as the ones from three-level models. When covariates are included in the HLMs, Zhu et al. (2012) found that in almost all situations the results will be nearly identical regardless of whether or not the classroom or middle level is omitted when designing or analyzing CRTs. However, one exception occurs when the middle-level variance is relatively large, such as in secondary schools where the proportion of classroom-level variance ranges from 0.293 to 0.376, the SE will be biased if the classroom level is omitted. In addition, when the study being planned has a markedly different cluster structure than the study that was used for planning purposes (e.g., the number of classrooms per school is halved), the MDESs from the two- and three-level analyses might yield quite comparable results. In these scenarios, adding student-level covariates that can explain a large proportion of the outcome variance (e.g., a pretest) can effectively eliminate any

potential problems. Therefore, educational researchers can utilize two-level analyses on three-level data without significant concern regarding the analysis of the main effects. However, it should be noted that omitting the middle level would make it impossible to investigate the classroom-level moderator or mediator effects. In addition, when only one classroom is sampled from each school, a two-level model with students nested within classrooms should be utilized to estimate the treatment effects and the school-level covariates can be added at the second level. When the randomization is at the classroom level (i.e., a multisite design), the classroom level should not be omitted from the analysis.

Non-compliance. During the implementation of the treatment, some participants may switch between the treatment and control groups. When some intervention participants do not comply with the random assignment, two options are available to estimate the intervention effects: the intention-to-treat (ITT) analysis, and the instrumental variable (IV) analysis (e.g., Angrist, 2006; Imbens & Rubin, 2015). We use the two-level design as an example to illustrate these methods. In particular, the ITT analysis compares the treatment and control groups according to the original random assignment, namely

Students Level:

$$Y_{ij} = \beta_{0j} + \mathbf{X_{ij}B_{1j}} + e_{ij},\ e_{ij} \sim N\left(0, \sigma^2_{|\mathbf{X}}\right), \tag{30}$$

School Level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} A_j + \mathbf{Z_j \Gamma_{02}} + u_{0j},\ u_{0j} \sim N\left(0, \tau_{|A,\mathbf{Z}}\right), \tag{31}$$

$$\mathbf{B_{1j}} = \mathbf{\Gamma_{10}}, \tag{32}$$

where A_j is a binary initial assignment indicator coded as one if a school is randomly assigned to the treatment group and zero otherwise. The parameter of interest—γ_{01}—provides an unbiased

estimate of the causal effect of participants' intention to treat but not the treatment effect because it ignores the non-compliance.

The IV analysis has been widely utilized in experimental studies to deal with non-compliance, and identify the causal effect of the intervention when the non-compliance is not random (e.g., Angrist, 2006; Konstantopoulos et al., 2016). Specifically, it uses the initial random assignment (e.g., A_j) as the IV of the actual receipt of the treatment (e.g., D_j) to estimate the causal effect of the treatment for compliers, which is usually called the local average treatment effect (LATE; Angrist, Imbens, & Rubin, 1996). Educational researchers can utilize the IV approach through the widely applied two-stage least-square (2SLS) estimation combined with cluster robust SEs or the 2SLS random effect estimator to deal with the nested data structure of CRTs (see Baltagi, 2013 or Wooldridge, 2010 for technical details). The IV approach can also be applied to moderation and mediation analyses (e.g., Dippel, Ferrara, & Heblich, 2020; Wooldridge, 2010).

Heterogeneous Variance. We typically assume the homogeneity of variances when utilizing HLMs to evaluate the main, moderator, and mediator effects. Wooldridge (2019) provided an overview of testing for heteroskedasticity (e.g., White test; White, 1997) for ordinary least square regressions. These methods can be applied to HLMs by diagnosing the level-1 residuals and the estimated random effects at the second and third levels. When the homogeneity assumption is violated, the SEs and significance test will be biased. Educational researchers can address this issue by modeling systematic heteroskedasticity, which requires correctly specifying the heteroskedasticity function form (Wooldridge, 2019). An alternative and easier way is to compute the heteroskedasticity-robust SE that has been implemented in major statistical software such as R and Stata. It should be noted that traditional heteroskedasticity-

robust SE requires at least 50 clusters to be effective (Huang, Wiedermann, & Zhang, 2023). Huang et al. assessed a specific robust SE estimator (i.e., the CR2 estimator; Bell & McCaffrey, 2002) using a Monte Carlo simulation and found it effective when the number of clusters was smaller than 50. They also provided R syntax to apply the CR2 SEs (available from https://github.com/flh3/CR2).

Blocking. Educational research can group similar clusters (e.g., schools) together and then randomize them into treatment and control conditions within each block to improve the precision of the estimated ATE. Although prior literature on blocking experiments focused on designs without nested data structures, their conclusions apply to CRTs. Specifically, based on the discussion from Pashley and Miratrix (2022), we recommend keeping the proportion of treated clusters similar across blocks. In addition, blocking is an excellent strategy to increase efficiency with respect to the effects of school-level moderators because it guarantees the balance of these moderators between treatment and control clusters (Dong, Kelcey, & Spybrook, 2018). It should be noted that, with a limited number of clusters, blocking multiple school-level moderators is challenging. Thus, following Athey and Imbens (2017), we suggest blocking a few key moderators of interest and making sure there are at least four clusters (e.g., two in the treatment and two in the control) per block.

External Validity. When examining the validity of a CRT, one needs not only to assess whether it yields an unbiased estimate of the ATE within the study sample (i.e., internal validity) but also on the population of interest (i.e., external validity). CRTs in education are typically conducted with a nonrandomly selected sample of schools that are recruited based on logistics factors such as convenience, the anticipated extent of implementation fidelity, the anticipated extent of responsiveness to the intervention, and cost (Olsen et al., 2013; Tipton & Olsen, 2018).

Prior studies (e.g., Olsen et al., 2013; Bell et al., 2016) have shown that nonrandomly selected samples yield biased impact estimates for the broader population of interest and thus are a vital threat to the external validity of CRTs. For example, Bell et al. (2016) indicated that for the Reading First program that was conducted in a nonrandomly selected sample of school districts from 11 educational impact evaluations, the impact estimates would be substantially biased downward with 0.10 standard deviations lower than the impact in the broader population.

Methods and tools have been developed to help educational researchers to improve the external validity of CRTs. A strategy to improve the representativeness of sites (e.g., schools or districts) is to stratify the population based on a set of characteristics that might moderate the impacts of the intervention and then systematically or randomly select sites from each stratum (Olsen & Orr, 2016; Tipton et al., 2014; see Tipton & Olsen, 2018 for details on steps to conduct stratifications). Researchers can then assess the representativeness of the selected sites by examining the proportion of the population being represented (i.e., coverage), calculating the standardized mean difference in the estimated program impacts between the distribution of the selected sites and the distribution of the target population, and/or doing both of these through calculating a generalizability index (Tipton & Olsen, 2018). Tipton and Miller (2016) provided a free web tool (www.thegeneralizer.org) for stratifying the population of interest and assessing the representativeness of the selected sample. It should be noted that, to successfully generalize results from CRTs, educational researchers must identify all potential moderators and the distribution of these moderators for the population of interest. Specifically, all school-level moderators (including student-level aggregates) must be known to generalize the results from a CRT to the school population of interest, and all student-level and school-level moderators must be known to generalize the results from a CRT to the student population of interest. The lack of

inclusion of important moderator variables may lead to erroneous conclusions. That is, if not all effect moderators have been measured, there is no guarantee that adjustments will necessarily 'improve' the generalizability of the effect. In fact, such adjustments might actually worsen it. Attempting to generalize a CRT effect based solely on the observed effect moderators could be more misleading than helpful for decision-makers.

To further reduce bias when estimating the ATE for the population, researchers can consider using propensity score methods to post-stratify or subclassify the sample (Tipton et al., 2014), regression modeling (Tipton & Olsen, 2018), or bounding approaches (Chan, 2017). Tipton and Olsen (2022) provided a comprehensive guide on steps and techniques to design and implement impact studies in education to make the findings more generalizable to the study's target population.

Cost-Effectiveness Analysis. While causal evaluations of educational interventions' effects are essential, it is equally important to examine the cost of achieving these effects for more sound policy decision-making (Bowden, 2017; Monk, 1995; Ross et al., 2007). Evaluations without a credible cost analysis might lead to misleading judgments and decisions. For example, when the research team evaluates the computer-assisted tutoring program, they might find no significant difference in the effectiveness compared to a traditional in-person tutoring program. Still, schools, districts, and parents might see the CAL tutoring program as an attractive alternative to traditional face-to-face teacher professional development programs if it could be delivered at a much lower cost. Therefore, it is recommended that the research team incorporates a cost study and explore the cost-effectiveness of their program.

The CRT that evaluates both the cost and effectiveness of an intervention is commonly named the cluster randomized cost-effectiveness trial (CRCET; Li et al., 2022). It links the cost

of implementing an intervention to its effect and thus helps researchers and policymakers adjudicate the cost-effectiveness of an intervention. Similar to the effectiveness data (e.g., test scores), the cost data also have a nested structure, and lower-level (e.g., student-level) cost is correlated within the same clusters (e.g., classrooms or schools). Therefore, recent studies (e.g., Li et al., 2020) recommend using HLMs to account for the dependencies of the cost data and evaluate CRCETs. They also developed a user-friendly tool—PowerUp!-CEA (Li et al., 2023)—to support applied researchers in performing statistical power analysis for CRCETs.

Conclusion

This primer introduces the statistical methods in the design and analysis of CRTs that have been widely utilized to evaluate the causal effects of educational interventions. Specifically, we define causal estimands based on the potential outcome framework and utilize HLM to account for the dependency of the intervention participants within the same clusters. We demonstrate methods and tools for sample size planning and statistical power analysis. Additionally, we discuss common challenges and potential solutions in the design and analysis phases, including the effects of omitting one level of clustering, non-compliance, heterogeneous variance, blocking, threats to external validity, and cost-effectiveness of the intervention.

CRTs in education commonly have two- or three-level of clustering, and thus, two- and three-level HLMs are introduced to estimate the main effects (e.g., ATEs), moderator effects, and mediator effects. When the middle-level (e.g., classroom-level) information is not available, educational researchers can use two-level models to analyze data from three-level CRTs in most scenarios. It is also recommended to include student-level covariates that could explain a large proportion of the outcome variance at the middle level (e.g., pretest scores) when the middle-level information is missing.

One key consideration when designing a CRT is sample size planning through power analysis to guarantee a good enough chance to detect the effect of interest when it exists. We introduced statistical power analysis methods based on HLMs and demonstrated the power and MDES computation using PowerUp! tools. In general, the power of a CRT is determined by the sample sizes at all levels, effect size, ICCs, and the proportion of outcome variance explained by covariates at all levels. For both the two- and three-level designs, the sample size at the cluster level (e.g., schools) has a larger impact on power than the sample sizes at the lower levels (e.g., classes and students) holding other factors fixed (Konstantopoulos, 2017; Konstantopoulos, Li & Zhang, 2023; Li & Konstantopoulos, 2017; Raudenbush, 1997). We recommend educational researchers incorporate covariates into their HLM analyses to decrease the SEs of ATEs and increase power. Although many prior studies have provided reference values of the design parameters for the U.S., there is very limited information regarding these design parameters for Asian countries. To design CRTs with adequate power, one direction of future research is to estimate design parameters using large-scale datasets that include Asian countries or educational interventions conducted in Asia.

It is not rare that some CRT participants do not comply with the random assignment in practice, and we recommend educational researchers employ the IV approach to estimate the treatment effects for compliers or LATEs. Prior studies also discussed the methods of obtaining the range of ATE using the IV approach and bounds analysis (e.g., Athey & Imbens, 2018). Besides the effectiveness, educational researchers need to consider the cost and cost-effectiveness of an intervention for a comprehensive evaluation and solid decision-making. Recent developments in cost-effectiveness analysis and experimental designs have provided methods and guidelines for designing cost studies within CRTs using the ingredients method,

computing power for CRCETs, and performing cost-effectiveness analysis using HLMs (e.g., Bowden, 2023; Li et al., 2022).

HLMs usually perform well in analyzing nested data when the number of clusters is relatively large (e.g., > 40). However, CRTs in education sometimes only have fewer clusters because of logistic or financial restrictions. Prior studies have shown that HLMs can be applied to CRTs with about 20 to 40 clusters and suggested small-sample corrections for CRTs with 10 to 20 clusters (Bell et al., 2014; Kenward & Roger, 2009). Other studies provided alternative methods (e.g., generalized estimating equations, cluster bootstrapping, Bayesian methods, etc.) to analyze CRTs with a small number of clusters (e.g., Gelman, 2006; Huang, 2016; Morel et al., 2003). For example, Huang and Li (2022) found that the use of OLS regression together with the bias-reduced linearization (BRL) cluster robust SE (Bell & McCaffrey, 2002) and empirically based degrees of freedom yields unbiased results with acceptable type I error and power. They also developed R and Stata packages to implement this method (Huang & Zhang, 2022).

This primer introduces the moderator analysis to evaluate the treatment effect heterogeneity (TEH). Recent developments in statistics and econometrics proposed to use machine learning (ML) methods to explore the TEH by estimating the conditional average treatment (CATE). Compared to traditional interaction analysis, the ML methods have some advantages, such as allowing selecting moderators from a potentially large number of covariates and identifying the causal TEH. However, these methods usually assume the study participants are independent and thus cannot easily be applied to CRTs. Therefore, one direction of future research is to evaluate the performance of the ML methods for data with nested structures in exploring TEH. Another way of examining TEH is via quantile regression that examines the treatment effects across the distribution of the outcome (e.g., Konstantopoulos et al., 2019).

The current primer focuses on continuous outcomes (e.g., test scores), while binary outcomes (e.g., whether or not a student graduates from high school) are also frequently used in CRTs (e.g., Ding et al., 2021). Multilevel logistic regression or linear probability models can be used to estimate the treatment effect for CRTs with binary outcomes (see Raudenbush & Bryk, 2002; Wooldridge, 2010). Besides the model-based methods (e.g., HLMs) discussed in this primer, design-based methods and tools are also available for the design and analysis of CRTs. For example, the software RCT-YES (Schochet, 2015) uses a non-parametric design-based approach that does not require assumptions on the distributions of potential outcomes to evaluate the effectiveness of CRTs (Schochet, 2016). It can estimate ATE, moderator effects, and compiler average causal effects with valid SEs for a wide range of single- and multilevel-level designs. In particular, for CRTs, the design-based non-parametric methods require fewer clusters because the analysis can be conducted using data on cluster-level averages rather than individual-level data (Schochet, 2016).

This primer assumes no missing data in the design and analysis of CRTs. However, attrition or dropout is almost always expected in practice. Educational researchers can use modern missing data techniques such as multiple imputations (Little & Rubin, 2019) to address the missing values and then estimate the effects of interests using HLMs, as discussed above. Enders (2023) provided an overview of recent developments in missing data methodologies over the past two decades. He particularly discussed the methods of handling missing data for multilevel models, including joint model imputation, fully conditional specification, maximum likelihood estimation, Bayesian estimation and multiple imputation, and fixed effect imputation. Enders (2023) also introduced the current software, such as *Blimp* (Keller & Enders, 2021) and R package *mdmb* (Grund, Lüdtke, & Robitzsch, 2021).

This primer focused on cluster designs, while multisite designs are also very popular in education, where for example, students are nested within classrooms nested within schools and the treatment could be at the student or classroom level. Similar to CRTs, HLMs are widely used to analyze data from multisite randomized trials (e.g., Li & Konstantopoulos, 2019; Raudenbush & Liu, 2000). Prior studies also discussed utilizing alternative methods (e.g., fixed effect models; Miratix, Weiss, & Henderson, 2021; Dong et al., 2021) to estimate the main and moderator. Educational researchers can still use PowerUp! and PowerUp!-Moderator to plan their multisite studies.

Declarations

Ethical approval: Because this study focused on statistical methods, it was exempted from Institutional Review Board approval.

Conflict of interest: The author(s) declare no competing interests.

References

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, *2*, 23-44. https://doi.org/10.1007/s11292-005-5126-x

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444-455. https://doi.org/10.2307/2291629

Baltagi, B. H.(2008). *Econometric analysis of panel data* (5th ed). Chichester, UK: Wiley.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173. https://doi.org/10.1037/0022-3514.51.6.1173

Bell, R., & McCaffrey, D. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28, 169–182.

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10*(1), 1–11. https://doi-org.lp.hscl.ufl.edu/10.1027/1614-2241/a000062

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318-335. https://doi.org/10.3102/0162373715617549

Bloom, H. S. (1995). Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs. *Evaluation Review*, *19*(5), 547–556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research* (MDRC Working Papers on Research Methodology). New York, NY: Manpower Demonstration Research Corporation. Available from http://www.mdrc.org/publications/437/full.pdf

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), 30–59. https://doi.org/10.3102/016237370729955

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2022). PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.1.0. [Software]. https://CRAN.R-project.org/package=PowerUpR

Chan, W. (2017). Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, *10*(3), 646-669. https://doi.org/10.1080/19345747.2016.1273412

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101. http://www.jstor.org/stable/20182143

Cox, K. & Kelcey, B. (2019). Robustness of Statistical Power in Group-Randomized Studies of Mediation Under an Optimal Sampling Framework. *Methodology, 15, 3, 106-118*

Ding, Y, Li, W., Li, X., Yang, J, & Ye, X. (2021). Heterogeneous major preference for extrinsic incentives: The effects of wage information on the gender gap in STEM major choice. *Research in Higher Education, 62*, 1113–1145. https://doi.org/10.1007/s11162-021-09636-w

Dippel, C., Ferrara, A., & Heblich, S. (2020). Causal mediation analysis in instrumental-variables regressions. *The Stata Journal*, *20*(3), 613–626. https://doi.org/10.1177/1536867X20953572

DiTraglia, F. J., García-Jimeno, C., O'Keeffe-O'Donovan, R., & Sánchez-Becerra, A. (2023). Identifying causal effects in experiments with spillovers and non-compliance. *Journal of Econometrics*.

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses of moderator effects in three-level cluster randomized trials. *Journal of Experimental Education, 86* (3), 489-514. https://doi.org/10.1080/00220973.2017.1315714

Dong, N., Kelcey, B., & Spybrook, J. (2021). Design Considerations in Multisite Randomized Trials Probing Moderated Treatment Effects. Journal of Educational and Behavioral Statistics, 46(5), 527–559. https://doi.org/10.3102/1076998620961492

Dong, N., Kelcey, B., Spybrook, J., & Maynard, R. A. (2017a). PowerUp!-Moderator: A tool for calculating statistical power and minimum detectable effect size of the

moderator effects in cluster randomized trials (Version 1.08) [Software]. Available from http://www.causalevaluation.org/

Dong, N., Kelcey, B., Spybrook, J., & Maynard, R. A. (2017b). PowerUp!-Mediator: A tool for calculating statistical power for causally-defined mediation in cluster randomized trials. (Beta Version 1.0) [Software]. Available from http://www.causalevaluation.org/

Dong, N., & Maynard, R. (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. https://doi.org/10.1080/19345747.2012.673143

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for panning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review, 40*(4), 334-377. doi: 10.1177/0193841X16671283

Dong, N., Spybrook, J., Kelcey, B., & Bulus, M. (2021). Power analyses for moderator effects with (non)random slopes in cluster randomized trials. *Methodology, 17* (2), 92-110. doi: https://doi.org/10.5964/meth.4003

Donner, A., Klar, N., & Klar, N. S. (2000). *Design and analysis of cluster randomization trials in health research* (Vol. 27). London: Arn.

Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. Advance online publication:https://dx.doi.org/10.1037/met0000563

Fairchild, A. J., & McDaniel, H. L. (2017). Best (but oft-forgotten) practices: mediation analysis. *American Journal of Clinical Nutrition*, 105(6), 1259-1271.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*, 515–534.

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: a flexible sequential modeling approach. *Behavior Research Methods*, *53*, 2631-2649.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Hayes, R. J., & Moulton, L. H. (2017). *Cluster randomized trials*. CRC press.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hedges, L. V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265-275. https://doi.org/10.1080/00131881.2018.1493350

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(6), 546–582. https://doi.org/10.1177/0193841X14554212

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education*, *84*(1), 175-196. https://doi.org/10.1080/00220973.2014.952397

Huang, F. L., Wiedermann, W., & Zhang, B. (2022). Accounting for heteroskedasticity resulting from between-group differences in multilevel models. *Multivariate Behavioral Research*, 1-21.

Huang, F., & Zhang, B. (2022). CR2: Compute cluster robust standard errors with degrees of freedom adjustments. https://cran.r-project.org/web/packages/CR2/

Huang, F. L., Zhang, B., & Li, X. (2022). Using robust standard errors for the analysis of binary outcomes with a small number of clusters. *Journal of Research on Educational Effectiveness*, 1-33. https://doi.org/10.1080/19345747.2022.2100301

IES RFA (2023). Institute of Education Sciences Education Research Grants Program Request for Applications 84.305A. Available from https://ies.ed.gov/funding/pdf/2024_84305A.pdf

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Experimental Power for Indirect Effects in Group-randomized Studies with Group-level Mediators. *Multivariate Behavioral Research, 52, 6, 699-719.*

Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017). Statistical Power for Causally-defined Indirect Effects in Group-randomized Trials with Individual-level Mediators. *Journal of Educational and Behavioral Statistics, 42, 5, 499-530.*

Kelcey, B., Spybrook, J., & Dong, N. (2019). Sample size planning in cluster-randomized interventions probing multilevel mediation. *Prevention Science, 20* (3), 707-418. doi: 10.1007/s11121-018-0921-6

Kelcey, B., Xie, Y., Spybrook, J., & Dong, N. (2021). Power and sample size determination for multilevel mediation in three-level cluster-randomized trials. *Multivariate behavioral research*, *56*(3), 496-513.

Kelcey, B., Spybrook, J., Dong, N., & Bai, F. (2020). Cross-Level Mediation in School-Randomized Studies of Teacher Development: Experimental Design and Power. *Journal of Research on Educational Effectiveness, 13*(3), 459-487.

Keller, B. T., & Lenders, C. K. (2021). Blimp user's guide (Version 3). https://www.appliedmissingdata.com/blimp

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583-2595. https://doi.org/10.1016/j.csda.2008.12.013

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*(1), 66-88. https://doi.org/10.1080/19345740701692522

Konstantopoulos, S., Miller, S., van der Ploeg, A. & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness, 9(SI),* 188-208. https://doi.org/10.1080/19345747.2015.1116031

Konstantopoulos, S., Li, W., Miller, S., & van der Ploeg, A. (2019). Using quantile regression to estimate intervention effects beyond the mean. *Educational and Psychological Measurement*, *79(5)*, 883-910.

Konstantopoulos, S., Li, W., & Zhang, B. (2023). Statistical power in cross-sectional multilevel experiments in education. In Stemmler, M., Wiedermann, W., &

Huang, F. (eds.), Dependent data in social sciences research: Forms, issues, and methods of analysis (2nd ed.), Springer.

Li, W., Dong, N. & Maynard, R. (2020). Power analysis for two-level multisite randomized cost-effectiveness trials. *Journal of Educational and Behavioral Statistics,* 45(6), 690–718. https://doi.org/10.3102/1076998620911916

Li, W., Dong, N., Maynard, R., Spybrook, J. & Kelcey, B. (2022). Experimental design and statistical power for cluster randomized cost-effectiveness trials. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2022.2142177

Li, W., & Konstantopoulos, S. (2017). Power analysis for models of change in cluster randomized designs. *Educational and Psychological Measurement, 77,* 119-142.

Li, W., & Konstantopoulos, S. (2019). Power computations for polynomial change in block randomized designs. *Journal of Experimental Education*, *87(4)*, 575-595.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593-614.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, *85*(3), 353-394. https://doi.org/10.3102/0034654314553127

Miratrix, L. W., Weiss, M. J., & Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, *14*(1), 270-308. https://doi.org/10.1080/19345747.2020.1831115

Mo, D., Zhang, L., Wang, J., Huang, W., Shi, Y., Boswell, M., & Rozelle, S. (2014). *The persistence of gains in learning from computer assisted learning (CAL): Evidence from a randomized experiment in rural schools in Shaanxi province in China*. (Working Paper No. 268). https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/Persistence_of_Gains_in_Learning_from_CAL.pdf

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate behavioral research*, *39*(1), 129-149.

Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, *45*, 395–409.

Olsen, R. B., & Orr, L. L. (2016). On the "where" of social experiments: Selecting more representative samples to inform policy. *New Directions for Evaluation*, *2016*(152), 61-71. https://doi.org/10.1002/ev.20207

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*(1), 107-121. https://doi.org/10.1002/pam.21660

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological methods*, *2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213. https://doi.org/10.1037/1082-989x.5.2.199

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. https://doi.org/10.3102/0162373707299460

Rhoads, C. and Li, Y. (2022). Causal Inference in Multi-level Settings. In O'Connell, A., McCoach D.B. & Bell, B.A. (Eds.) Multilevel Modeling Methods with Introductory and Advanced Applications.

Schochet, P. Z. (2015). Statistical theory for the RCT-YES software: Design-based causal inference for RCTs (NCEE 2015–4011). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Schochet, P.Z. (2016). RCT-YES software: User's Manual.

Sim, M., Kim, S.-Y., & Suh, Y. (2022). Sample Size Requirements for Simple and Complex Mediation Models. *Educational and Psychological Measurement*, 82(1), 76-106. https://doi.org/10.1177/00131644211003261

Spybrook, J., Bloom, H., Gongdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal Design plus empirical evidence: Documentation for the "optimal design" software. Retrieved April 20, 2021, from http://hlmsoft.net/od/od-manual-20111016-v300.pdf, http://hlmsoft.net/od/

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41*(6), 605-627. https://doi.org/10.3102/1076998616655442

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open, 2*(1), 1–15.

https://doi.org/10.1177/2332858415625975

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educational Evaluation and Policy Analysis*, *42*(3), 354-374. https://doi.org/10.3102/0162373720929018

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114-135. https://doi.org/10.1080/19345747.2013.831154

Tipton, E., & Miller, K. (2016). *The Generalizer: A webtool for improving the generalizability of results from experiments*. http://www.thegeneralizer.org

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524. https://doi.org/10.3102/0013189X18781522

Tipton, E., & Olsen, R. B. (2022). *Enhancing the Generalizability of Impact Studies in Education*. (NCEE 2022-003). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, *2*(4), 457-468.

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, *237*(1), 105-237.

Vo, T. T., Superchi, C., Boutron, I., & Vansteelandt, S. (2020). The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. *Journal of clinical epidemiology*, 117, 78-88.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817-838.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and Analyzing Studies That Randomize Schools to Estimate Intervention Effects on Student Academic Outcomes Without Classroom-Level Information. Educational Evaluation and Policy Analysis, 34(1), 45–68. https://doi.org/10.3102/0162373711423786

Table 1. List of parameters used for power and MDES computations

Parameters	Meaning of the parameters and computation formulas
δ_2	Effect size for two-level design; $\delta_2 = \frac{\gamma_{01}}{\sqrt{\tau^2+\sigma^2}}$
δ_3	Effect size for three-level design; $\delta_3 = \frac{\gamma_{001}}{\sqrt{\omega^2+\tau^2+\sigma^2}}$
ρ	Intra-class correlation coefficient (ICC) for two-level designs; $\rho = \frac{\tau^2}{\tau^2+\sigma^2}$
ρ_2	Intra-class correlation coefficients (ICCs) at the second level for three-level designs; $\rho_2 = \frac{\tau^2}{\omega^2+\tau^2+\sigma^2}$
ρ_3	Intra-class correlation coefficients (ICCs) at the third level for three-level designs; ; $\rho_3 = \frac{\omega^2}{\omega^2+\tau^2+\sigma^2}$
R_1^2	Proportion of level-1 variance explained by level-1 covariates; $R_1^2 = 1 - \frac{\sigma^2_{\mid X}}{\sigma^2}$
R_2^2	Proportion of level-2 variance explained by level-2 covariates; $R_2^2 = 1 - \frac{\tau^2_{\mid w}}{\tau^2}$
R_3^2	Proportion of level-3 variance explained by level-3 covariates; $R_3^2 = 1 - \frac{\omega^2_{\mid Z}}{\omega^2}$

Note: To be conservative, we usually assume the treatment indicator does not explain any proportion of the outcome variance.

Table 2. Summary of HLMs, Non-centrality Parameter, MDES, and Degrees of Freedom for the Analysis of Main Effects

Model Name	Models	Standardized Noncentrality (λ) Parameter and MDES	Degrees of Freedom
Two-Level Unconditional Model	Level-1: $Y_{ij} = \beta_{0j} + e_{ij},\ e_{ij} \sim N(0, \sigma^2)$ Level-2: $\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + u_{0j},\ u_{0j} \sim N(0, \tau^2)$	$\lambda_{2u} = \delta_2 \sqrt{\frac{P(1-P)Jn}{1+(n-1)\rho}}$ $MDES_{2u} = M_{J-2} \sqrt{\frac{\rho}{P(1-P)J} + \frac{(1-\rho)}{P(1-P)Jn}}$	*J-2*
Two-Level Conditional Model	Level-1: $Y_{ij} = \beta_{0j} + \mathbf{B_{1j}X_{ij}} + e_{ij},\ e_{ij} \sim N(0, \sigma^2_{\mid \mathbf{X}})$ Level-2: $\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \mathbf{\Gamma_{02}W_j} + u_{0j},\ u_{0j} \sim N(0, \tau^2_{\mid T,\mathrm{W}})$ $B_{1j} = \Gamma_{01}$	$\lambda_{2c} = \delta_2 \sqrt{\frac{P(1-P)Jn}{n\rho(1-R_2^2) + (1-\rho)(1-R_1^2)}}$ $MDES_{2C} = M_{J-g-2} \sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}}$	*J-2-g*
Three-Level Unconditional Model	Level-1: $Y_{ijk} = \beta_{0jk} + e_{ijk},\ e_{ij} \sim N(0, \sigma^2)$ Level-2: $\beta_{0jk} = \pi_{00k} + r_{0jk},\ r_{0jk} \sim N(0, \tau^2)$ Level-3: $\pi_{01k} = \gamma_{000} + \gamma_{001} T_k + u_{0k},\ u_{0j} \sim N(0, \omega^2)$	$\lambda_{3u} = \delta_3 \sqrt{\frac{P(1-P)JKn}{1+(n-1)\rho_2 + (Jn-1)\rho_3}}$ $MDES_{3C} = M_{K-2} \sqrt{\frac{\rho_3}{P(1-P)K} + \frac{\rho_2}{P(1-P)JK} + \frac{1-\rho_2-\rho_3}{P(1-P)JKn}}$	*K-2*
Three-Level Conditional Model	Level-1: $Y_{ijk} = \beta_{0jk} + \mathbf{B_{1jk}X_{ijk}} + e_{ijk},\ e_{ij} \sim N(0, \sigma^2_{\mid \mathbf{X}})$ Level-2: $\beta_{0jk} = \pi_{00k} + \mathbf{\Pi_{01k}W_{jk}} + r_{0jk},\ r_{0jk} \sim N(0, \tau^2_{\mid \mathbf{w}})$ $\mathbf{B_{1jk}} = \mathbf{\Pi_{10k}}$ Level-3: $\pi_{01k} = \gamma_{000} + \gamma_{001} T_k + \mathbf{\Gamma_{02}Z_k} + u_{0k},\ u_{0j} \sim N(0, \omega^2_{\mid T,\mathbf{Z}})$ $\mathbf{\Pi_{01k}} = \mathbf{\Gamma_{010}}$ $\mathbf{\Pi_{10k}} = \mathbf{\Gamma_{100}}$	$\lambda_{3c} = \delta_3 \sqrt{\frac{P(1-P)JKn}{Jn\rho_3(1-R_3^2) + n\rho_2(1-R_2^2) + (1-\rho_2-\rho_3)(1-R_1^2)}}$ $MDES_{3C} = M_{K-q-2} \sqrt{\frac{\rho_3(1-R_3^2)}{P(1-P)K} + \frac{\rho_2(1-R_2^2)}{P(1-P)JK} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)JKn}}$	*K-2-q*

Note: (1) *P* represents the proportion of clusters (e.g., schools) assigned to the treatment group, n represents the sample size at level 1, *J* represents the sample size at level 2, and *K* represents the sample size at level 3. (2) *g* represents the number of covariates at level 2 and *q* represents the number of covariates at level 3.

Table 3. Demonstration of MDES Computation for Two-level Cluster Designs using PowerUp!

Model 3.1: MDES Calculator for Two-Level Cluster Random Assignment Design (CRA2_2)— Treatment at Level 2		
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)	0.23	Proportion of variance in outcome that is between clusters
P	0.50	Proportion of Level 2 units randomized to treatment: $J_T / (J_T + J_C)$
R_1^2	0.50	Proportion of variance in Level 1 outcomes explained by Level 1 covariates
R_2^2	0.50	Proportion of variance in Level 2 outcome explained by Level 2 covariates
g*	1	Number of Level 2 covariates
n (Average Cluster Size)	100	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
J (Sample Size [# of Clusters])	40	Number of Level 2 units
M (Multiplier)	2.88	Computed from T_1 and T_2
T_1 (Precision)	2.03	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.85	Determined from given power level
MDES	**0.314**	Minimum Detectable Effect Size

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

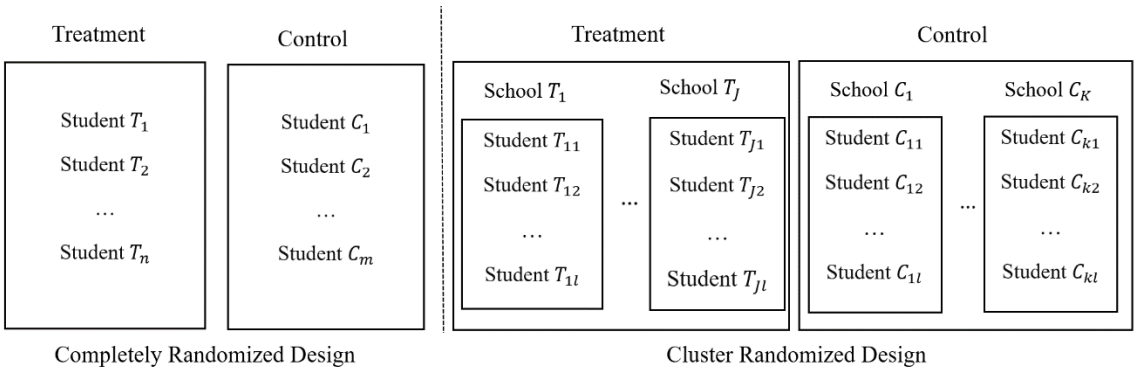

Figure 1. Illustration of Completely Randomized Design and Cluster Randomized Design

Model

```
You have requested statistical power for: Three-level Cluster-randomized Trial
```

Result

```
Statistical power:
--------------------------------------
 0.843
--------------------------------------
Degrees of freedom: 48
Standardized standard error: 0.083
Type I error rate: 0.05
Type II error rate: 0.157
Two-tailed test: TRUE
```

Figure 2. Example of PoweUpR Shinny App Output: Power Computation for Three-level Cluster Designs

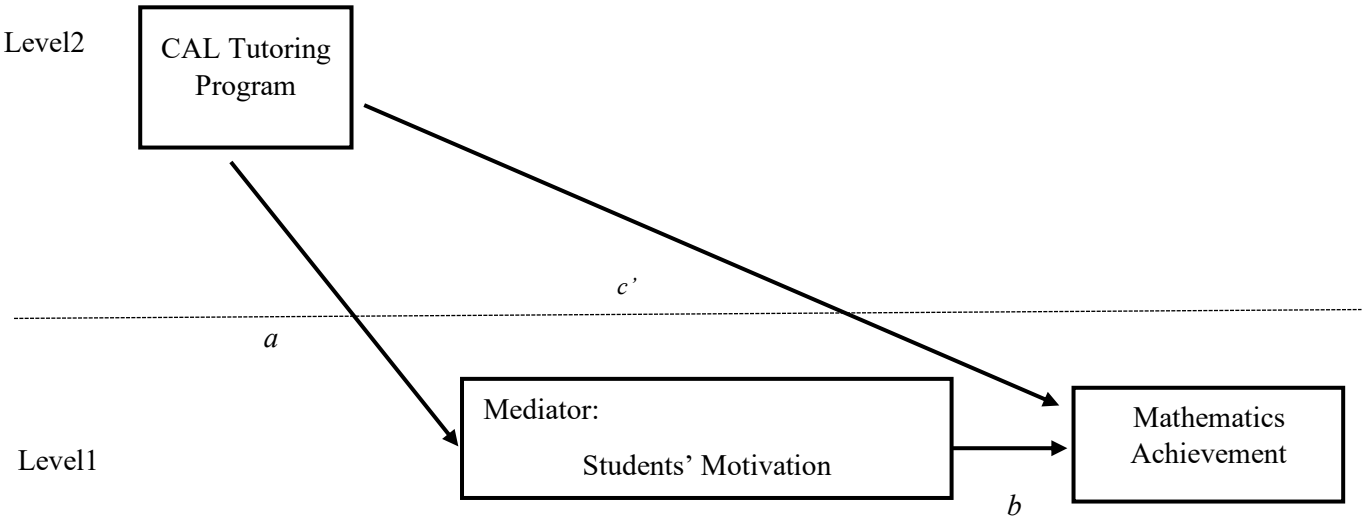

Figure 3. A 2-1-1 mediation model