ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);179-201

Estimation and Standardization of Variance Parameters for Planning Cluster-Randomized Trials: A Short Guide for Researchers

Metin BULUŞ* Sakine GÖÇER ŞAHİN**

Abstract
A review of literature covering the past decade indicates a shortage of cluster-randomized trials (CRTs) in education and psychology in Turkey, the gold standard that is capable of producing high-quality evidence for high-stake decision making when individual randomization is not feasible. Scarcity of CRTs is not only detrimental to collective knowledge on the effectiveness of interventions but also hinders efficient design of such studies as prior information is at best incomplete or unavailable. In this illustration, we demonstrate how to estimate variance parameters from existing data and transform them into standardized forms so that they can be used in planning sufficiently powered CRTs. The illustration uses publicly available software and guides researchers step by step via introducing statistical models, defining parameters, relating them to notations in statistical models and power formulas, and estimating variance parameters. Finally, we provide example statistical power and minimum required sample size calculations.

Key Words: cluster-randomized trials, variance estimation, statistical power analysis, minimum required sample size.

INTRODUCTION

Cluster randomized trials (CRTs) are experimental designs where subjects are not assigned to treatment conditions independently but rather as a group. There has been an increasing interest in CRTs over the past decade in educational research (Spybrook, Shi, & Kelcey, 2016). Merely using "CRT" as a searching keyword, more than 1000 articles related to CRTs are found in educational research area in the academic journals on the Web of Science database. Although CRTs are not as efficient as individual-randomized trials, the nature of an intervention may warrant assignment of clusters (groups of individuals) to treatment conditions. There are a couple of reasons for this. First, it may be more viable to implement an intervention at the cluster level. Second, using existing clusters can be highly beneficial in terms of cost reduction and implementation convenience. Third, it may not be ethical to deprive some subjects from the intervention within the same organization. For example, providing some students with a promising intervention while excluding others from the study could be considered an unfair practice in education. Furthermore, CRTs can reduce the risk of treatment contamination that might occur if individuals in the same organization were to be randomized to treatment conditions.

However, compared to individual-randomized trials, CRTs are more complicated to design, need more participants to obtain similar statistical power, and anticipated statistical analyses are more complicated (Hayes & Moulton, 2017). Statistical methods that ignore clustering might produce misleading results, because they assume that all subjects, regardless of which clusters they come from, provide independent observations. In education settings, the assumption of independent observations is often violated as a result of contextual effects. For example, observations may not be independent

* Research Associate / Lecturer, Adıyaman University, Adıyaman, TURKEY, e-mail: bulusmetin@gmail.com, ORCID ID: orcid.org/0000-0003-4348-6322
** Postdoctoral Researcher, University of Wisconsin-Madison, WI, USA, e-mail: sgocersahin@gmail.com, ORCID ID: orcid.org/0000-0002-6914-354X

To cite this article:
Buluş, M. & Göçer Şahin, S. (2019). Estimation and standardization of variance parameters for planning cluster-randomized trials: a short guide for researchers. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 179-201. DOI: 10.21031/epod.530642

Received: 22.02.2019
Accepted: 13.06.2019

from each other because students in the same classroom have an experienced teacher or collaboration among them is encouraged. Similarly, students and teachers within the same school share resources such as library or laboratory that differ from other schools, which may have similar contextual effect. Applying methods that ignore clustering (e.g. ordinary least squares) in such cases can prompt confidence intervals that are excessively narrow and yield p-values that are very small (Bland, 2004). In the case of experimental designs, narrower confidence intervals and smaller p-values can misguide researchers as they may indicate significant differences when, in fact, there is actually none.

There are different ways of addressing clustering depending on statistical methodology and sampling scheme. One solution is to make inferences based on cluster-robust standard errors (e.g. Cameron & Miller, 2015). If results pertain to a specific subpopulation consisting of a few clusters and not to be generalized, another alternative is to include cluster membership as fixed effects in the statistical model along with the treatment indicator. Nonetheless, applying Hierarchical Linear Models (HLM, Raudenbush & Bryk, 2002) is more common in education. Even if researchers can use cluster-robust standard errors, or depending on the sampling scheme, use fixed effects estimation method, it is not straightforward to decompose variance to within and between clusters, a strategy we will use throughout this guide to estimate and standardize variance parameters. Therefore, in parallel with studies in education effectiveness research we adopt HLM formulation.

By the same token, when planning studies that have similar nesting structure (student within classroom within schools) contextual effects should be taken into consideration, as power analysis procedures rely on the standard error of the estimate. There are various studies that have derived approximate standard error formulas with which a researcher can estimate power rate ahead of an experimental study (a priori power analysis) given sample size and other characteristics (e.g., Bloom, 1995; Bloom, 2006; Bloom, Bos, & Lee, 1999; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Konstantopoulos, 2009a, 2009b).

Despite the increasing trend in the use of CRTs across many education systems and countries around the world, our review of literature in the past decade indicates a shortage in educational and psychological research in Turkey. Also, statistical power analysis in existing studies are either absent or have not considered nesting structure of the sample. We examined 174 experimental studies in education field published in Turkish journals on the Ulakbim Tubitak Journal Park Database to see whether they report power analysis procedure to determine effective sample size. Although none of the experiments utilized CRT, none of the authors reported power analysis procedure either. As a result, in these papers, results mostly suffer from small sample size where the experiment possibly could not detect a significant treatment effect when in fact there was.

One particular issue with a priori power analysis is that variance parameters used in the approximate formulas are not known. Other parameters needed for power calculations either have commonly accepted standards or does not need estimation or require extensive methodological expertise. For example, standard practice in educational research is to keep power rate at 80%, have type I error rate of 5%, and to conduct two-tailed hypothesis testing of the treatment effect (Dong & Maynard, 2013). Moreover, sample size information (e.g., the average number of students per school) can be obtained from administrative records or calculated via descriptive statistics.

While there is an emerging body of literature reporting standardized variance parameters from existing data (e.g., Hedberg & Hedges, 2014; Hedberg, 2016; Hedges & Hedberg, 2013; Spybrook, Westine, & Taylor, 2016; Westine, 2016; Westine, Spybrook, & Taylor, 2014; Zopluoglu, 2012), the majority of which focuses on K-12 academic outcomes within the United States, results may not apply to other subjects, grades, or geographical areas. Variance parameters are often sample and subject specific and should be obtained either from prior research in the literature or empirical data, preferably as close as possible to the geographical area of interest, and as similar as possible to the subject under scrutiny. Thus, estimation and standardization of variance parameters from earlier research of the same kind become an indispensable tool to researchers, especially where there is little or no prior information.

Purpose of the Study

The purpose of this study is to guide researchers in education and psychology toward planning efficient CRTs in light of little or no prior knowledge. Specifically, the study aims to provide readers with a short tutorial on estimating variance parameters from existing data using HLM, standardizing them in terms of well-known variance parameters such intra-class correlation coefficients and R-squared values, and using standardized parameters in statistical power and minimum required sample size calculations for planning CRTs.

METHOD

We provide models for two- and three-level CRTs in HLM and mixed-model forms and define parameters as in Dong and Maynard (2013). We also illustrate how to estimate treatment effect and obtain variance parameters via `lme4` (Bates, Maechler, Bolker, & Walker, 2015) library in the R environment (R Core Team, 2019). Finally, we use estimated variance parameters (unstandardized) to calculate some of the standardized design parameters (i.e., intra-class correlation coefficients and R-squared values) and use them in statistical power analysis via `PowerUpR` (Bulus, Dong, Kelcey, & Spybrook, 2019). In most instances using two libraries in the R environment will be sufficient to analyze and plan CRTs, however, depending on the complexity of the task, researchers can use any other preferred software or platform.

Ideally, results from a CRT should be informative with respect to variation in the outcome, explanatory power of covariates, and the treatment effect, which can be obtained via several statistical models. Minimally sufficient models that can inform researchers in both planning and analysis of CRTs are null and full models. Null model (also known as unconditional model) can be used to get a sense of unconditional variation in the outcome (i.e., dependent variable), whereas full model can be used to estimate both the treatment effect and conditional variation in the outcome. Null and full models for two- and three-level CRTs are described below.

Two-level CRTs

Null Model to Estimate Unconditional Variation

The following unconditional model can be used to obtain variance parameters σ^2 and τ^2 as defined below, which will be used to calculate standardized variance parameters along with parameters from full model.

HLM formulation:

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$

Mixed model formulation:

$$Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij}$$

where r_{ij} and μ_{0j} are level 1 and level 2 residuals, following normal distributions as $r_{ij} \sim N(0, \sigma^2)$ and $\mu_{0j} \sim N(0, \tau^2)$, respectively. Thus, σ^2 and τ^2 are variances in the outcome between level 1 and level 2 units, respectively. Y_{ij} is level 1 outcome of interest for subject i in cluster j, β_{0j} is level 1 intercept (in this case, the mean of subjects in cluster j), γ_{00} is level 2 intercept (in this case, the mean of all subjects in all clusters - grand mean).

Full Model to Estimate Treatment Effect and Conditional Variation

The following model can be used to obtain variance parameters $\sigma^2_{|X}$ and $\tau^2_{|W}$ as defined below, which are used to calculate standardized variance parameters along with parameters from unconditional model.

HLM formulation:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \delta T_j + \gamma_{01}W_j + \mu_{0j}$

$\beta_{1j} = \gamma_{10}$

Mixed model formulation:

$$Y_{ij} = \gamma_{00} + \delta T_j + \gamma_{01}W_j + \gamma_{10}X_{ij} + \mu_{0j} + r_{ij}$$

where r_{ij} and μ_{0j} are level 1 and level 2 conditional residuals, following normal distributions as $r_{ij} \sim N(0, \sigma^2_{|X})$ and $\mu_{0j} \sim N(0, \tau^2_{|W})$, respectively. Thus, $\sigma^2_{|X}$ and $\tau^2_{|W}$ are conditional variances in the outcome between level 1 and level 2 units, respectively. Y_{ij} is level 1 outcome of interest for subject i in cluster j, X_{ij} is level 1 covariate for subject i in cluster j, T_j is treatment condition (1 if cluster j assigned to the treatment, 0 if not) and W_j is level 2 covariate for cluster j, β_{0j} is level 1 intercept, γ_{00} is level 2 intercept, δ is the treatment effect, β_{1j} or γ_{10} is regression weight for level 1 covariate X_{ij}, γ_{01} is regression weight for level 2 covariate W_j.

We can calculate standardized variance parameters based on unstandardized variance parameters from unconditional and full models. $\rho = \tau^2/(\tau^2 + \sigma^2)$ represents proportion of variance in the outcome between level 2 units (also referred to as intra-class correlation coefficient in the literature), $R^2_1 = 1 - \sigma^2_{|X}/\sigma^2$ is proportion of variance in the outcome explained by level 1 covariates, $R^2_2 = 1 - \tau^2_{|W}/\tau^2$ is proportion of variance in the outcome explained by level 2 covariates. The treatment effect can also be standardized in the form of Cohen's d as $\delta^* = \delta/\sqrt{\tau^2 + \sigma^2}$, hereafter often referred to as effect size.

In the full model, we can get an estimate for the treatment effect and the associated t statistics. The hypothesis of "there is a treatment effect" is tested against the null hypothesis of "there is no treatment effect". By comparing the t statistics from the full model to the critical t value given Type I error rate (α, probability of detecting treatment effect when in fact there is none in the underlying population), we can inspect whether results can be explained beyond chance factor. Similarly, knowing t statistics, we can have an idea about Type II error rate (β, probability of detecting no treatment effect when in fact there is an effect in the underlying population). In practice we are interested in the probability of detecting a treatment effect when in fact there is an effect in the underlying population, and that is statistical power ($1 - \beta$). To calculate statistical power, we can use t statistics after an experiment, although it may not be useful, as the experiment has already been completed. However, we can plan for an experiment such that sample size will likely produce adequate statistical power had it been repeated many times. To calculate statistical power prior to an experiment, we need some information from earlier studies; an estimate of what would be a meaningful treatment effect (often set as 0.20 or 0.25 in education, but may be increased if there is sufficient evidence that earlier interventions produced large treatment effects) and its standard error.

Standard Error Formula under Balanced Sample Size and Homogenous Variance

Assuming that level 1 variances are equal across J number of level 2 units, and level 1 sample sizes are balanced (e.g., n number of level 1 units per level 2 unit), standardized variance takes the form

$$Var(\delta^*) = \frac{\rho(1-R_2^2)}{p(1-p)J} + \frac{(1-\rho)(1-R_1^2)}{p(1-p)nJ}$$

Standard error of the treatment effect is $SE(\delta^*) = \sqrt{Var(\delta^*)}$, and if we know δ^* and $SE(\delta^*)$, we can calculate t statistics with which statistical power can be calculated. $\delta^*/SE(\delta^*)$ follows t distribution with $J-g-2$ degrees of freedom where g is number of covariates added at level 2 (Bloom, 2006, p. 17; Dong & Maynard, 2013, p. 51). Statistical power $(1-\beta)$ for two-tailed hypothesis testing can be calculated as

$$1-\beta = P\left(t_{df}(\lambda) > t_{df,1-\alpha/2}(0)\right) + P\left(t_{df}(\lambda) < t_{df,\alpha/2}(0)\right)$$

where $df = J-g-2$ for the two-level CRT, $t_{df,\alpha/2}(0)$ is the statistic associated with central t distribution with degrees of freedom df and probability $\alpha/2$, $t_{df}(\lambda)$ is the statistic associated with non-central t distribution with non-centrality parameter $\lambda = \delta^*/SE(\delta^*)$, degrees of freedom df, and α and β are Type I and Type II error rates (see, Hedges & Rhoads, 2010; Moerbeek & Safarkhani, 2018). In what follows we will demonstrate how to estimate variance parameters and how to calculate parameters needed in $Var(\delta^*)$ formula.

Estimation and Standardization of Treatment Effect and Variance Components

If not pre-installed, `lme4` and `PowerUpR` libraries should be installed in the R environment using `install.packages(c("lme4", "PowerUpR"))` command. They can be loaded into the current R session using `library(lme4)` and `library(PowerUpR)` commands.

In order to demonstrate variance estimation procedure in R, considering education settings, we simulate a simple two-level CRT data named `CRT2` which has 2,000 students across 100 schools (20 students per school). The data include five variables; school identification numbers (`schid`), a level 1 outcome variable (`outcome`), a level 2 treatment variable (`treatment`), a level 1 covariate (`covx`), and a level 2 covariate (`covw`). Number of level 1 or level 2 covariates will not change analysis strategy very much. Outcome is continuous and can be considered as any of the achievement indicator for a particular subject – such as mathematics, science, or reading scores. The treatment can be any intervention that aims at increasing student achievement scores such as a science, technology, engineering, and mathematics (STEM) program. Level 1 and level 2 covariates can be student pretest score and average school-level pretest score. First a few lines of the simulated data is printed below. Each school has a unique identification number (`schid`). Since schools are assigned to treatment conditions, the same school identification numbers will have the same values for treatment variable (`treatment`). Level 1 (students) and level 2 (schools) covariates (`covx` and `covw`) follows standard normal distributions, and outcome (`outcome`) is a linear function of these covariates with some level 1 and level 2 noise added (See data generation mechanism in Appendix A). From this point forward, R scripts are within shaded boxes. Along with code chunks, comments begin with `## --` and outputs begin with `##`.

```
head(CRT2)

##   schid treatment    outcome        covx      covw
## 1     1         0 -0.7145407 -0.37560287 0.2533185
## 2     1         0  0.2411899 -0.56187636 0.2533185
## 3     1         0 -0.8423327 -0.34391723 0.2533185
## 4     1         0 -0.9780591  0.09049665 0.2533185
## 5     1         0  3.2965023  1.59850877 0.2533185
## 6     1         0  1.7267023 -0.08856511 0.2533185
```

First, we estimate variance parameters for unconditional model to calculate the intra-class correlation coefficient. The output includes variance for two random effects indicating variation in the outcome that is between school means (`tau2`) and that is between students (`sigma2`). Sum of the two is roughly same as variance of the outcome. Thus, proportion of variance in the outcome that is between schools, also known as intra-class correlation (`rho2`), can be calculated.

```
## -- install.packages(c("lme4", "PowerUpR"))
library(lme4) # for estimation
library(PowerUpR) # for power analysis

## -- null model (unconditional model)
null.model <- lmer(outcome ~ (1 | schid), data = CRT2)
print(VarCorr(null.model), comp = "Variance")

##  Groups   Name        Variance
##  schid    (Intercept) 1.2253
##  Residual             1.9601

## -- variance parameters
tau2 <- 1.2253
sigma2 <- 1.9601


## -- intra-class correlation coefficient
rho2 <- tau2 / (tau2 + sigma2)
round(rho2, 2)

## [1] 0.38
```

Next, we estimate variance parameters for the full model to calculate R-squared values along with variance parameters from unconditional model. The output, again, includes variances for two random effects indicating conditional variation in the outcome that is between schools (`tau2w`) and students (`sigma2x`) beyond what is explained by level 2 and level 1 predictors. As some of the variation between schools and students are explained by level 2 and level 1 predictors respectively, note that variance components are reduced compared to the null model. Using proportion of reduction in the variance for level 2 and level 1, we can calculate R-squared values for each (`r21` and `r22`).

```
## -- full model
full.model <- lmer(outcome ~ treatment + covx + covw + (1 | schid),
data = CRT2)
print(VarCorr(full.model), comp = "Variance")

##  Groups   Name        Variance
##  schid    (Intercept) 0.85332
##  Residual             0.98335

## -- variance parameters
tau2w <- 0.8533
sigma2x <- 0.9834

## -- R-squared values for level 1 and level 2
r21 <- 1 - (sigma2x / sigma2)
r22 <- 1 - (tau2w / tau2)

round(r21, 2)

## [1] 0.5

round(r22, 2)

## [1] 0.3
```

We can also extract and standardize the treatment effect (`delta`) by the variance of the outcome in the form of Cohen's *d* (`es`). In this way, the effect is comparable to previous literature, can be compared to in future studies, and also be used in statistical power analysis procedures, if needed.

```
## -- treatment effect
coef(summary(full.model))["treatment",]

##   Estimate Std. Error    t value
##  0.9849094  0.1930537  5.1017374

delta <- 0.9849
es <- delta / sqrt(sigma2 + tau2)
round(es, 2)

## [1] 0.55
```

Statistical Power and Minimum Required Sample Size Calculations

Before we find statistical power and minimum required sample size, there are a few things to clarify. Earlier, we estimated and standardized variance parameters so that we can use them in power analysis procedures, however, there are other parameters needed, most of which are either have commonly accepted standards or known (or can be obtained via simple procedures that does not require methodological expertise). In education research, it is common to find power for an effect size (`es`) of 0.20 or 0.25, have a Type I error rate (`alpha`) of .05, and assume a two-tailed (`two.tailed`) hypothesis testing. Other way around, when the interest is in finding minimum required sample size, additionally, the power rate is assumed to be 80%. Furthermore, assigning half of the schools to treatment group (`p`) produces optimal power rate or optimal minimum required sample size (note that $p(1-p)$ in the denominator of standard error formula is maximum when $p = .50$). In our case, we

know there are 20 students per school (`n`), and 100 schools (`J`) in total. Now we can calculate statistical power as

```
## -- power analysis
design <- power.cra2r2(es = .20, alpha = .05, two.tailed = TRUE,
                       rho2 = .38, r21 = .50, g2 = 1, r22 = .30,
                       p = .50, n = 20, J = 100)

##
## Statistical power:
## ---------------------------------------
##  0.463
## ---------------------------------------
## Degrees of freedom: 97
## Standardized standard error: 0.106
## Type I error rate: 0.05
## Type II error rate: 0.537
## Two-tailed test: TRUE
```

where, in addition to parameters defined earlier, `g2` is the number of covariates added at level 2. Parameters obtained from the data produce a power rate of 46.3%, which means if we repeat this experiment for a large number of times, we will detect a statistically significant treatment effect 46.3% of the time, if in fact there is an effect in the underlying population. This is under recommended benchmark power rate of 80% in power analysis literature. In other words, this is worse than flipping a coin in order to decide whether or not an intervention would be effective. Figure 1 demonstrates how far we are from the benchmark power rate. By visual inspection, it seems a sample consisting of somewhere between 200 to 250 schools is capable of producing results with 80% power rate.

```
plot(design, ypar = "power", locate = TRUE, xlim = c(50, 250))
```

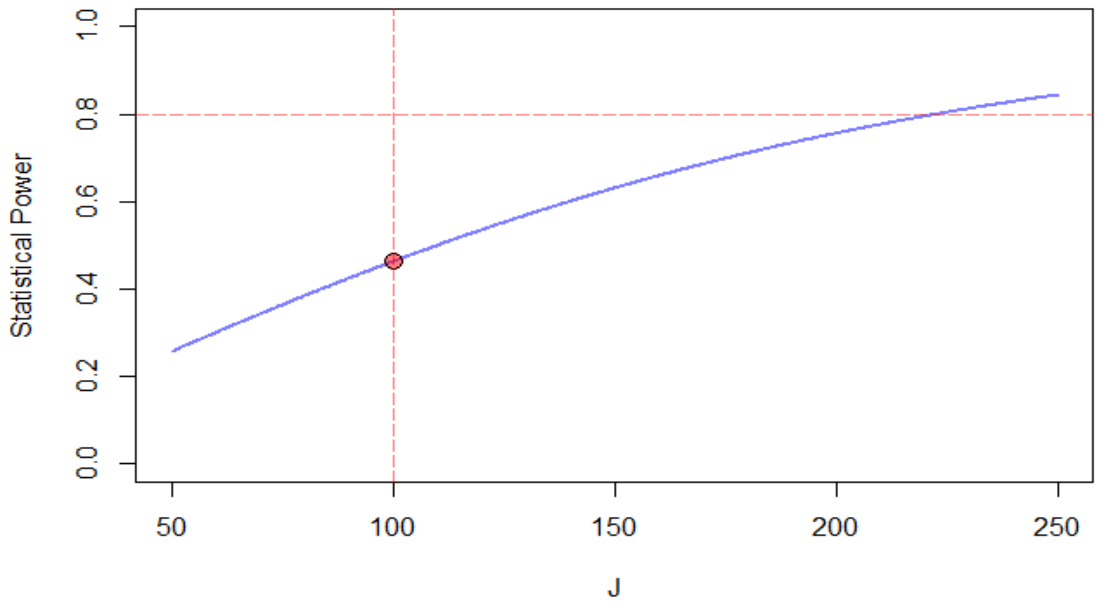

Figure 1. Statistical Power as a Function of Number of Schools for Two-level CRT Example

Precise number of schools to detect an effect size of 0.20 with 80% power rate can be found via calculating minimum required number of schools in PowerUpR (script below) or PowerUp! (Figure 2) as

```
# -- minimum required sample size
mrss.cra2r2(power = .80, es = .20, alpha = .05, two.tailed = TRUE,
            rho2 = .38, r21 = .50, g2 = 1, r22 = .30,
            p = .50, n = 20)

## J = 223
```

Model 3.1: Sample Size Calculator for 2-Level Cluster Random Assignment Design (CRA2_2_)— Treatment at Level 2		
Assumptions		**Comments**
MRES = MDES	0.20	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)	0.38	Proportion of variance in outcome that is between clusters
n (Average Cluster Size)	20	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
Sample Retention Rate: Level 2 units	100%	Proportion of Level 2 units retained in analysis sample
Sample Retention Rate: Level 1 units	100%	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 Level 1 covariates
R_2^2	0.300	Proportion of variance in Level 2 outcome explained by Level 2 covariates
g*	1	Number of Level 2 covariates
Priori-M (Multiplier)	2.81	Computed from Priori-T1 and Priori-T2
M (Multiplier)	2.81	Automatically computed
J (Sample Size [Clusters #])	**223**	Number of clusters needed for given MRES

RUN

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

Figure 2. Minimum Required Number of Schools for Two-level CRT Example

With a sample similar to what we have in terms of average of number of students per school ($n = 20$), intra-class correlation coefficient ($\rho = .38$), explanatory power of covariates at level 1 ($R_1^2 = .50$), and at level 2 ($R_2^2 = .30$), we need at least 223 schools to detect an effect size of 0.20 with a power rate of 80% and type I error rate of 5% for a two-tailed hypothesis testing of the treatment effect.

Explanatory Power of Covariates

Researchers often have control over sample size to increase power rate prior to implementing a two-level CRT. However, in some cases, sampling more units is not feasible or induces prohibitive cost. In this case, explanatory power of covariates for a level can be increased via collecting more information, which in turn improves the power rate. The question naturally comes to mind is whether to collect more information on level 1 or level 2 units. To address this question, we demonstrate to what extent changes in R_1^2 or R_2^2 lead to changes in variance for treatment effect via taking first derivative of $Var(\delta^*)$ with respect to R_1^2 or R_2^2. What becomes apparent is that changes in $Var(\delta^*)$ occur in the opposite direction with changes in R_1^2 or R_2^2 (note negative signs). This means if we increase R_1^2 or R_2^2 this will reduce $Var(\delta^*)$, which in turn improves power rate.

$$\frac{\partial Var(\delta^*)}{\partial R_2^2} = -\frac{\rho}{p(1-p)J}$$

$$\frac{\partial Var(\delta^*)}{\partial R_1^2} = -\frac{(1-\rho)}{p(1-p)nJ}$$

Due to limited resources, researchers may favor collecting information on a level that reduces $Var(\delta^*)$ comparably more. In this case, increasing R_2^2 reduces the variance $(\rho n)/(1-\rho)$ times more compared to the reduction induced by increasing R_1^2 by the same amount (obtained from the ratio of the two derivatives). Therefore, focusing on increasing explanatory power of covariates at level 2 is a more efficient strategy.

For example, for the two-level CRT example, increasing R_2^2 from .40 to .50 (.10 increment) reduces variance from 0.01126 to 0.00974 (a reduction of 0.00152), which, in turn, increases power rate from 46.3% to 51.9%. However, increasing R_1^2 from .30 to .40 (.10 increment) marginally reduces variance from 0.01126 to 0.011136 (a reduction of 0.000124), which, in turn, increases power rate marginally from 46.3% to 46.7%. The ratio of variance reductions is precisely what one would obtain if they use $(\rho n)/(1-\rho)$ formula, which is 12.26. This means increasing R_2^2 by .10 reduces variance 12.26 times more compared to the variance reduction induced by increasing R_1^2 by the same amount.

Three-level CRTs

Null Model to Estimate Unconditional Variation

The following unconditional model can be used to obtain variance parameters σ^2, τ_2^2 and τ_3^2 as defined below, which will be used to calculate standardized variance parameters along with parameters from the full model.

HLM formulation:

Level 1: $Y_{ijk} = \beta_{0jk} + r_{ijk}$

Level 2: $\beta_{0jk} = \gamma_{00k} + \mu_{0jk}$

Level 3: $\gamma_{00k} = \xi_{000} + \varsigma_{00k}$,

Mixed model formulation:

$$Y_{ijk} = \xi_{000} + \varsigma_{00k} + \mu_{0jk} + r_{ijk}$$

where r_{ijk}, μ_{0jk}, and ς_{00k} are level 1, level 2, and level 3 residuals, following normal distributions as $r_{ijk} \sim N(0, \sigma^2)$, $\mu_{0jk} \sim N(0, \tau_2^2)$, and $\varsigma_{00k} \sim N(0, \tau_3^2)$, respectively. Thus, σ^2, τ_2^2 and τ_3^2 are variances in the outcome between level 1, level 2 and level 3 units, respectively. β_{0jk} is level 1 intercept (in this case, mean of subjects in sub-cluster j and cluster k), γ_{00k} is level 2 intercept (in this case, mean of subjects in all sub-clusters in cluster k), ξ_{000} is level 3 intercept (in this case, mean of all subjects in all sub-clusters in all clusters - grand mean).

Full Model to Estimate Treatment Effect and Conditional Variation

The following model can be used to obtain variance parameters $\sigma^2_{|X}$, $\tau^2_{2|W}$ and $\tau^2_{3|V}$ as defined below, which are used to calculate standardized variance parameters along with parameters from the unconditional model.
HLM formulation:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{ijk} + r_{ijk}$

Level 2: $\beta_{0jk} = \gamma_{00k} + \gamma_{01k}W_{jk} + \mu_{0jk}$
$\beta_{1jk} = \gamma_{10k}$
Level 3: $\gamma_{00k} = \xi_{000} + \delta T_k + \xi_{001}V_k + \varsigma_{00\mathrm{k}}$
$\gamma_{01k} = \xi_{010}$
$\gamma_{10k} = \xi_{100}$

Mixed model formulation:

$$Y_{ijk} = \xi_{000} + \delta T_k + \xi_{001}\mathrm{V_k} + \xi_{010}W_{jk} + \xi_{100}X_{ijk} + \varsigma_{00\mathrm{k}} + \mu_{0jk} + r_{ijk}$$

where r_{ijk}, μ_{0jk}, and $\varsigma_{00\mathrm{k}}$ are conditional residuals following normal distributions as $r_{ijk}\sim N(0, \sigma^2_{|X})$, $\mu_{0jk}\sim N(0, \tau^2_{2|W})$, and $\varsigma_{00\mathrm{k}}\sim N(0, \tau^2_{3|W})$, respectively. Thus, $\sigma^2_{|X}$, $\tau^2_{2|W}$ and $\tau^2_{3|V}$ are residual variances at level 1, level 2 and level 3, respectively, which are not accounted for by the full model. Y_{ijk} is level 1 outcome of interest for subject i in sub-cluster j which is in cluster k, X_{ijk} is level 1 covariate for individual i in sub-cluster j which is in cluster k, W_{jk} is level 2 covariate for sub-cluster j in cluster j, T_k is treatment condition (1 if cluster k assigned to treatment, 0 if not), and V_k is level 3 covariate. β_{0jk}, γ_{00k}, and ξ_{000} are level 1, level 2 and level 3 intercepts, respectively. δ is the treatment effect, β_{1jk} or γ_{10k} or ξ_{100} is regression weight for level 1 covariate X_{ijk}, γ_{01k} or ξ_{010} is regression weight for level 2 covariate W_{jk}, and ξ_{001} is regression weight for level 3 covariate V_k.

Similar to two-level CRT case, we can calculate standardized variance parameters based on unstandardized variance parameters from unconditional and full models. $\rho_2 = \tau_2^2/(\tau_3^2 + \tau_2^2 + \sigma^2)$ and represents proportion of variance in the outcome between level 2 units, $\rho_3 = \tau_3^2/(\tau_3^2 + \tau_2^2 + \sigma^2)$ and represents proportion of variance in the outcome between level 3 units, $R_1^2 = 1 - \sigma^2_{|X}/\sigma^2$ and is proportion of variance in the outcome explained by level 1 covariates, $R_2^2 = 1 - \tau^2_{2|W}/\tau_2^2$ and is proportion of variance in the outcome explained by level 2 covariates, and $R_3^2 = 1 - \tau^2_{3|V}/\tau_3^2$ and is proportion of variance in the outcome explained by level 3 covariates. The treatment effect can be standardized in the form of Cohen's *d* as $\delta^* = \delta/\sqrt{\tau_3^2 + \tau_2^2 + \sigma^2}$.

Standard Error Formula under Balanced Sample Size and Homogenous Variance

Assuming balanced sample sizes, that is, n number of level 1 units per level 2 unit, J number of level 2 units per level 3 unit, and also assuming variance within each level 2 and level 3 unit is same across JK number of level 2 units and K number of level 3 units, standardized standard error takes the form

$$Var(\delta^*) = \frac{\rho_3(1-R_3^2)}{p(1-p)K} + \frac{\rho_2(1-R_2^2)}{p(1-p)JK} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{p(1-p)nJK}$$

Similar to two-level CRT, standard error of the treatment effect is $SE(\delta^*) = \sqrt{Var(\delta^*)}$. If we know δ^* and $SE(\delta^*)$ we can calculate t statistics, and therefore statistical power can be calculated. $\delta^*/SE(\delta^*)$ follows t distribution with $K - g_3 - 2$ degrees of freedom where g_3 is number of

covariates added at level 3 (Dong & Maynard, 2013, p. 52). Statistical power can be calculated as in the two-level CRT case.

Estimation and Standardization of Treatment Effect and Variance Components

Similar to two-level CRT case, considering education settings, we simulated a simple three-level CRT data named `CRT3` which has 6000 students across 300 classrooms in 100 schools (20 students per classroom and 3 classrooms per school). The data includes seven variables; school identification numbers (`schid`), classroom identification numbers (`clsid`), a level 1 outcome variable (`outcome`), a level 3 treatment variable (`treatment`), a level 1 covariate (`covx`), a level 2 covariate (`covw`), and a level 3 covariate (`covv`). First few lines of the simulated data are printed below. Each school and classroom have unique identification numbers (`schid` and `clsid`). Since schools are assigned to treatment conditions, the same school and classrooms therein will have the same values for treatment variable (`treatment`). Level 1 (students) and level 2 (classrooms), and level 3 (schools) covariates (`covx`, `covw`, and `covv`) follow standard normal distributions, and outcome (`outcome`) is a linear function of these covariates with some level 1, level 2, and level 3 noise added (See data generation mechanism in Appendix A).

```
head(CRT3)

##   schid clsid treatment    outcome       covx       covw      covv
## 1     1     1         0  3.0263592  0.5622673 -0.3756029 0.2533185
## 2     1     1         0  1.7124732 -0.0974125 -0.3756029 0.2533185
## 3     1     1         0  1.0353372  1.0164552 -0.3756029 0.2533185
## 4     1     1         0 -0.8436311 -1.1561674 -0.3756029 0.2533185
## 5     1     1         0  1.7452900  2.3208602 -0.3756029 0.2533185
## 6     1     1         0  0.6092003 -0.6035312 -0.3756029 0.2533185
```

As in two-level CRT case, first we estimate variance parameters for unconditional model to calculate intra-class correlation coefficients. The output includes variance for three random effects indicating variation in the outcome that is between school means (`tau23`), between classroom means (`tau22`) and that is between students (`sigma2`). Sum of the three is roughly same as variance of the outcome. Thus, proportion of variance in the outcome that is between schools and classrooms can be calculated (`rho3` and `rho2`).

```
## -- null model (unconditional model)
null.model <- lmer(outcome ~ (1 | schid) + (1 | clsid), data = CRT3)
print(VarCorr(null.model), comp = "Variance")

##  Groups   Name        Variance
##  clsid    (Intercept) 1.2593
##  schid    (Intercept) 0.9969
##  Residual             1.6160

## -- variance parameters
tau23 <- 0.9969
tau22 <- 1.2593
sigma2 <- 1.6160

## -- intra-class correlation coefficients for level 2 and level 3
rho2 <- tau22 / (tau23 + tau22 + sigma2)
rho3 <- tau23 / (tau23 + tau22 + sigma2)
round(rho2, 2)

## [1] 0.33

round(rho3, 2)

## [1] 0.26
```

The output for the full model, again, includes variance for three random effects indicating conditional variation in the outcome that is between schools (`tau23v`), classrooms (`tau22w`) and students (`sigma2x`) beyond what is explained by level 3, level 2 and level 1 predictors, respectively. As some of the variation between schools, between classrooms and between students are explained by level 3, level 2 and level 1 variables, using proportion of reduction in the variance for level 3, level 2 and level 1 we can calculate R-squared values for each (`r23`, `r22` and `r21`).

```
## --  full model
full.model <- lmer(outcome~ treatment + covx + covw + covv +
                (1 | schid) + (1 | clsid), data = CRT3)
print(VarCorr(full.model), comp = "Variance")

##  Groups   Name        Variance
##  clsid    (Intercept) 1.06824
##  schid    (Intercept) 0.71853
##  Residual             1.00901

## -- variance parameters
tau23v <- 0.7185
tau22w <- 1.0682
sigma2x <- 1.0090

## -- R-squared values for level 1, level 2 and level 3
r21 <- 1 - (sigma2x / sigma2)
r22 <- 1 - (tau22w / tau22)
r23 <- 1 - (tau23v / tau23)
round(r21, 2)

## [1] 0.38

round(r22, 2)

## [1] 0.15

round(r23, 2)

## [1] 0.28

## -- treatment effect
coef(summary(full.model))["treatment",]

##   Estimate Std. Error    t value
##  0.9323254  0.2124156  4.3891572

delta <- 0.9323
es <- delta / sqrt(sigma2 + tau22 + tau22)
round(es, 2)

## [1] 0.46
```

Statistical Power and Minimum Required Sample Size Calculations

Default parameters for power analysis are same as two-level CRT case. Different from two-level CRT case, there are 20 students per classroom (`n`), 3 classrooms per school (`J`), and 100 schools (`K`) in total. Now we can calculate statistical power as

```
## -- power analysis
design <- power.cra3r3(es = .20, alpha = .05, two.tailed = TRUE,
                       rho2 = .33, rho3 = .26,
                       r21 = .38, r22 = .15, g3 = 1, r23 = .28,
                       p = .50, n = 20, J = 3, K = 100)

##
## Statistical power:
## ---------------------------------------
##  0.458
## ---------------------------------------
## Degrees of freedom: 97
## Standardized standard error: 0.107
## Type I error rate: 0.05
## Type II error rate: 0.542
## Two-tailed test: TRUE
```

where, in addition to calculated parameters above, `g3` is number of covariates added at level 3. Parameters obtained from the data produce a power rate of 45.8%, which means if we repeat this experiment for a large number of times, we will detect a statistically significant treatment effect 45.8% of the time, if in fact there is a treatment effect in the underlying population. Figure 3 demonstrates how far we are from the benchmark power rate. By visual inspection, it seems a sample consisting of somewhere between 200 to 250 schools is capable of producing results with 80% power rate.

```
plot(design, ypar = "power", locate = TRUE, xlim = c(50, 250))
```

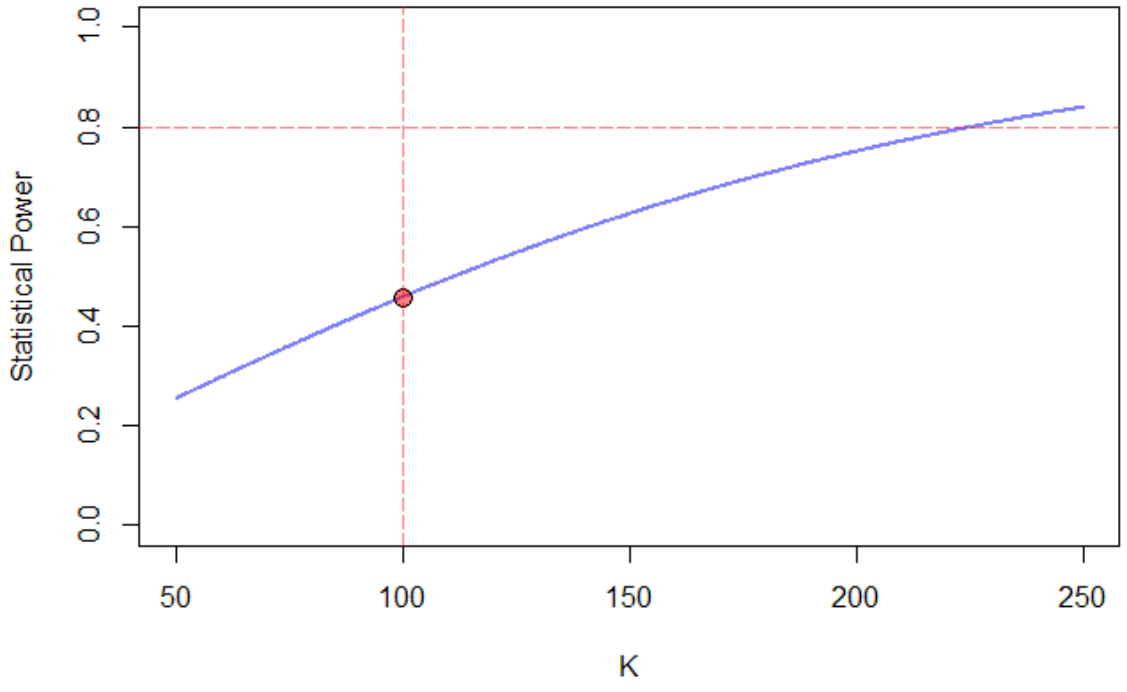

Figure 3. Statistical Power as a Function of Number of Schools for Three-level CRT Example

To find minimum required number of schools needed to detect an effect size of 0.20 with a power rate of 80% we can use PowerUpR (script below) or PowerUp! (Figure 4) as

```
# -- minimum required sample size
mrss.cra3r3(power = .80, es = .20, alpha = .05, two.tailed = TRUE,
            rho2 = .33, rho3 = .26,
            r21 = .38, r22 = .15, g3 = 1, r23 = .28,
            p = .50, n = 20, J = 3)

## K = 226
```

Model 3.2: Sample Size Calculator for 3-Level Cluster Random Assignment Designs (CRA3_3r)— Treatment at Level 3		
Assumptions		Comments
MRES = MDES	0.20	Minimum Relevant Effect Size = Minimum Detectable Effect Size
Alpha Level (α)	0.05	Probability of Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1 - probability of Type II error)
Rho_3 (ICC_3)	0.26	Proportion of variance in outcome between Level 3 units: V3/(V1+V2+V3)
Rho_2 (ICC_2)	0.33	Proportion of variance between Level 2 units: V2/(V1 + V2 + V3)
P	0.50	Proportion of Level-3 units randomized to treatment
R_1^2	0.38	Proportion of variance in Level 1 outcome explained by the Level 1 covariates
R_2^2	0.15	Proportion of variance in Level 2 outcome explained by the Level 2 covariates
R_3^2	0.28	Proportion of variance in Level 3 outcome explained by the Level 3 covariates
g_3*	1	Number of Level 3 covariates
n (Average Sample Size for Level 1)	20	Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)
J (Average Sample Size for Level 2)	3	Mean number of Level 2 units per Level 3 unit (harmonic mean recommended)
Priori-J (Sample Size [Clusters #])	226	
Priori-T_1 (for desired precision)	1.97	Computed from given alpha Level, two-tailed or one-tailed test
Priori-T_2 (for desired precision)	0.84	Computed from given power Level
Priori-M (Multiplier)	2.81	Computed from Priori-T1 and Priori-T2
M (Multiplier)	2.81	Automatically computed
K (Sample Size [# of Level 3 units])	**226**	Number of Level 3 clusters needed for given MDES.

RUN

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

Figure 4. Minimum Required Number of Schools for Three-level CRT Example

With a sample similar to what we have in terms of average number of students per classroom ($n = 20$), average number of classrooms per school ($J = 3$), intra-class correlation coefficients ($\rho_2 = .33$

and $\rho_3 = .26$), explanatory power of covariates at level 1 ($R_1^2 = .38$), level 2 ($R_2^2 = .15$), and at level 3 ($R_3^2 = .28$), power analysis result suggest that we need at least 226 schools to detect an effect size of 0.20 with a power rate of 80% and type I error rate of 5% for a two-tailed hypothesis testing of treatment effect.

Explanatory Power of Covariates

Due to the same reasons and similar to two-level CRT case, one should keep in mind that it is more efficient to increase explanatory power of covariates via including additional covariates at the third level. If we take first derivative of $Var(\delta^*)$ with respect to R_1^2, R_2^2, or R_3^2, what becomes apparent is that changes in $Var(\delta^*)$ occur in the opposite direction with changes in R_1^2, R_2^2, or R_3^2. This means increase in explanatory power for any of the R_1^2, R_2^2, or R_3^2 will reduce $Var(\delta^*)$, which improves the power rate.

$$\frac{\partial Var(\delta^*)}{\partial R_3^2} = -\frac{\rho_3}{p(1-p)K}$$

$$\frac{\partial Var(\delta^*)}{\partial R_2^2} = -\frac{\rho_2}{p(1-p)JK}$$

$$\frac{\partial Var(\delta^*)}{\partial R_1^2} = -\frac{(1-\rho_2-\rho_3)}{p(1-p)nJK}$$

Comparably, increasing R_3^2 reduces the variance $(\rho_3 J)/\rho_2$ times more compared to the reduction induced by increasing R_2^2 by the same amount, and $(\rho_3 nJ)/(1-\rho_2-\rho_3)$ times more compared to the reduction induced by increasing R_1^2. Therefore, focusing on increasing explanatory power of covariates at level 3 is a more efficient strategy.

For example, for the three-level CRT example, increasing R_3^2 from .28 to .38 (.10 increment) reduces variance from 0.011398 to 0.010357 (a reduction of 0.00104), which, in turn, increases power rate from 45.8% to 49.4%. Similarly, increasing R_2^2 from .15 to .25 (.10 increment) reduces variance from 0.011398 to 0.010957, which, in turn, increases power rate from 45.8% to 47.3%. The ratio of variance reductions is precisely what one would obtain if they use $(\rho_3 J)/\rho_2$ formula, which is 2.36. This means increasing R_3^2 by .10 reduces variance 2.36 times more compared to the variance reduction induced by increasing R_2^2 by the same amount. However, increasing R_1^2 from .48 to .58 (.10 increment) reduces variance marginally from 0.011398 to 0.011370, which, in turn, increases power rate marginally from 45.8% to 45.9%. Ratio of variance reductions is precisely what one would obtain if they use $(\rho_3 nJ)/(1-\rho_2-\rho_3)$ formula, which is 38. This means increasing R_3^2 by .10 reduces variance 38 times more compared to the variance reduction induced by increasing R_1^2 by the same amount.

DISCUSSION and CONCLUSION

In this tutorial, we demonstrated how to analyze and plan two- and three-level CRTs. We provided statistical models and estimated variance parameters to further use them in statistical power analysis procedures. Most of the power analysis programs require specification of standardized variance parameters. We also demonstrated how to standardize variance parameters into intra-class correlation coefficients and R-squared values. This guide will potentially assist researchers in their endeavors to plan two- and three-level CRTs with greater precision, thus, provide reliable results to evaluators, stakeholders and policy makers.

Statistical power calculations for two- and three-level CRTs can be conducted in any software program that allows standardized parameters as input (e.g., Optimal Design Plus, PowerUpR and PowerUp!).

Results from minimum required sample size (MRSS) calculations in PowerUp! and PowerUpR are compared to each other in nine slightly different designs (D1-D9 in Table B1) for two-level CRT, changing one parameter at a time. The same procedure is repeated for three-level CRT (D1-D12 in Table B2). Results indicate that MRSS calculations in both software programs are very much the same, rarely differ by one unit as a result of rounding difference in two different platforms.

We elaborated on the explanatory power of covariates and their relation to statistical power, demonstrated that collecting more information on higher level units and including them in statistical models as covariates improve power rate substantially. In contrast, covariates added at the individual level improve power rate only marginally. Thus, if there are financial and practical challenges to sampling more clusters, an alternative strategy would be focusing on improving explanatory power of covariates.

From the beginning of an intervention to the end, some clusters and individuals therein may refuse or discontinue participating, resulting in non-participation or attrition which deteriorates the power rate. Non-participation and attrition rates can also be obtained from prior research, for which minimum required sample size calculations can be adjusted accordingly. Thus, when analyzing existing data or reporting results, documenting non-participation and attrition rates will also help researchers to design CRTs with greater precision. One thing to keep in mind, in education context for example, is the fact that those students within schools cannot be oversampled while we can sample additional schools to adjust the sample size for non-participation or attrition.

There are some limitations to this guide. Although we demonstrated how to estimate variance parameters for CRTs, there might be other practical issues a researcher needs to deal with. For example, there might be missing data, outliers, or assumption of linearity may not hold. Researchers may also need to use weights, if they would like to plan for generalizable large-scale CRTs, and they have access to similar large-scale data sets. Such topics require an extensive treatment and are beyond the scope of this guide.

REFERENCES

Bland J. M. (2004). Cluster randomized trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*, *4*(21). DOI: https://doi.org/10.1186/1471-2288-4-21

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*(5), 547-556. DOI: 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: MDRC. Retrieved from DOI: https://www.mdrc.org/sites/default/files/full_533.pdf.

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445-469. DOI: https://doi.org/10.1177%2F0193841X9902300405

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2019). PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.0.4. DOI: https://CRAN.R-project.org/package=PowerUpR

Cameron, A. C., & Miller, d. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*, 317-372. DOI: https://doi.org/10.3368/jhr.50.2.317

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. DOI: https://doi.org/10.1080/19345747.2012.673143

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. DOI: https://doi.org/10.18637/jss.v067.i01

Hayes, R. J. & Moulton, L. H. (2017). *Cluster Randomized Trials* (2nd ed.). New York, NY: Chapman and Hall/CRC Press. DOI: https://doi.org/10.4324/9781315370286

Hedberg, E. C. (2016). Academic and behavioral design parameters for cluster randomized trials in kindergarten: an analysis of the Early Childhood Longitudinal Study 2011 Kindergarten Cohort (ECLS-K 2011). *Evaluation Review*, *40*(4), 279-313. DOI: https://doi.org/10.1177/0193841X16655657

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. DOI: https://doi.org/10.1177/0193841X14554212

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445-489. DOI: https://doi.org/10.1177/0193841X14529126

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. Retrieved from https://files.eric.ed.gov/fulltext/ED509387.pdf

Konstantopoulos, S. (2009a). Using power tables to compute statistical power in multilevel experimental designs. *Practical Assessment, Research & Evaluation*, *14*(10).

Konstantopoulos, S. (2009b). Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. *Evaluation Review*, *33*(4), 335-357. DOI: https://doi.org/10.1177/0193841X09337991

Moerbeek, M., & Safarkhani, M. (2018). The design of cluster randomized trials with random cross-classifications. *Journal of Educational and Behavioral Statistics*, *43*(2), 159-181. DOI: https://doi.org/10.3102/1076998617730303

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

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing [Computer software]. Vienna, Austria. Retrieved from https://www.R-project.org.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. International *Journal of Research & Method in Education*, *39*(3), 255-267. DOI: https://doi.org/10.1080/1743727X.2016.1150454

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1). DOI: https://doi.org/10.1177/2332858415625975

Westine, C. D. (2016). Finding Efficiency in the Design of Large Multisite Evaluations: Estimating Variances for Science Achievement Studies. *American Journal of Evaluation*, *37*(3), 311-325. DOI: https://doi.org/10.1177/1098214015624014

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490-519. DOI: https://doi.org/10.1177/0193841X14531584

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 242-278.

Appendix A
Data Generation Process

Data Generating Model for Two-level CRT

The statistical model to generate data for two-level CRT is same as the statistical model described in the main text. Here we provide only the mixed model formulation, which is

$$Y_{ij} = \gamma_{00} + \delta T_j + \gamma_{01} W_j + \gamma_{10} X_{ij} + \mu_{0j} + r_{ij}$$

where parameters are explained elsewhere in the main text. The following parameter values are used in the simulation, while considering 20 students per school (n) and 100 schools in total (J).

$$\gamma_{00} = 0$$
$$\delta = 1$$
$$T_j \sim BERN(0.50)$$
$$\gamma_{01} = 0.50$$
$$W_j \sim N(0,1)$$
$$\gamma_{10} = 1$$
$$X_{ij} \sim N(0,1)$$
$$\mu_{0j} \sim N(0,1)$$
$$r_{ij} \sim N(0,1)$$

```
set.seed(123) # for replication
delta <- 1
js <- 100
ns <- rep(20, js)
id <- as.factor(rep(1:js, ns))
tj <- rep(rbinom(js, 1, .50), ns)
wj <- rep(rnorm(js), ns)
uj <- rep(rnorm(js), ns)
xij <- rnorm(sum(ns))
rij <- rnorm(sum(ns))
yij <- delta * tj + 0.50 * wj + xij + uj + rij

CRT2 <- data.frame("schid" = id,
                   "treatment" = tj,
                   "outcome" = yij,
                   "covx" = xij,
                   "covw" = wj)
```

Data Generating Model for Three-level CRT

The mixed model formulation for three-level CRT is

$$Y_{ijk} = \xi_{000} + \delta T_k + \xi_{001} \mathrm{V_k} + \xi_{010} W_{jk} + \xi_{100} X_{ijk} + \varsigma_{00k} + \mu_{0jk} + r_{ijk}$$

where parameters are explained elsewhere in the main text. The following parameter values are used in the simulation, while considering 20 students per classroom (n), 3 classrooms per school (J), and 100 schools in total (K).

$$\xi_{000} = 0$$
$$\delta = 1$$
$$T_k \sim BERN(0.50)$$
$$\xi_{001} = 0.25$$
$$\mathrm{V_k} = N(0,1)$$
$$\xi_{010} = 0.50$$
$$W_{jk} \sim N(0,1)$$
$$\xi_{100} = 0.75$$

$$X_{ijk} \sim N(0,1)$$
$$\varsigma_{00k} \sim N(0,1)$$
$$\mu_{0jk} \sim N(0,1)$$
$$r_{ijk} \sim N(0,1)$$

```
set.seed(123) # for replication
delta <- 1
ks <- 100
js <- rep(3, ks)
ns <- rep(20, sum(js))

id3 <- as.factor(rep(rep(1:ks, js), ns))
id2 <- as.factor(rep(rep(1:sum(js), ns)))

tk <- rep(rep(rbinom(ks, 1, .50), js), ns)
vk <- rep(rep(rnorm(ks), js), ns)
sk <- rep(rep(rnorm(ks), js), ns)
wjk <- rep(rep(rnorm(sum(js)), ns))
ujk <- rep(rep(rnorm(sum(js)), ns))
xijk <- rnorm(sum(ns))
rijk <- rnorm(sum(ns))
yijk <- delta * tk + 0.25 * vk + 0.50 * wjk +  0.75 * xijk + sk + ujk +
rijk

CRT3 <- data.frame("schid" = id3,
                   "clsid" = id2,
                   "treatment" = tk,
                   "outcome" = yijk,
                   "covx" = xijk,
                   "covw" = wjk,
                   "covv" = vk)
```

Appendix B
PowerUpR and PowerUp! Comparisons

Table B1
Comparison for Two-level CRTs

Assumptions	**Base**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**
MRES = MDES	0.20	**0.40**	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Alpha Level (α)	0.05	0.05	**0.01**	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Two-tailed or One-tailed Test?	2	2	2	**1**	2	2	2	2	2	2
Power (1-β)	0.80	0.80	0.80	0.80	**0.20**	0.80	0.80	0.80	0.80	0.80
Rho (ICC)	0.40	0.40	0.40	0.40	0.40	**0.20**	0.40	0.40	0.40	0.40
n (Average Cluster Size)	20	20	20	20	20	20	**10**	20	20	20
P	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.30**	0.50	0.50
R_1^2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.20**	0.50
R_2^2	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	**0.50**
J (Sample Size [# of Level 2 units]) in ***PowerUp!***	**234**	**60**	**348**	**184**	**41**	**128**	**246**	**239**	**241**	**171**
J (Sample Size [# of Level 2 units]) in **PowerUpR**	**233**	**60**	**348**	**184**	**41**	**128**	**245**	**238**	**241**	**171**

Note. g (number of covariates added at level 2) is fixed at 1 for all nine designs.

Table B2
Comparison for Three-level CRTs

Assumptions	**Base**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**
MRES = MDES	0.20	**0.40**	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Alpha Level (α)	0.05	0.05	**0.01**	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Two-tailed or One-tailed Test?	2	2	2	**1**	2	2	2	2	2	2	2	2	2
Power (1-β)	0.80	0.80	0.80	0.80	**0.20**	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Rho_3 (ICC_3)	0.30	0.30	0.30	0.30	0.30	**0.15**	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Rho_2 (ICC_2)	0.30	0.30	0.30	0.30	0.30	0.30	**0.10**	0.30	0.30	0.30	0.30	0.30	0.30
P	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.30**	0.50	0.50	0.50	0.50	0.50
R_1^2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.30**	0.50	0.50	0.50	0.50
R_2^2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.40**	0.50	0.50	0.50
R_3^2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.70**	0.50	0.50
n (Average Sample Size for Level 1)	20	20	20	20	20	20	20	20	20	20	20	**10**	20
J (Average Sample Size for Level 2)	2	2	2	2	2	2	2	2	2	2	2	2	**3**
K (Sample Size [# of Level 3 units]) in ***PowerUp!***	**183**	**47**	**272**	**144**	**33**	**125**	**145**	**217**	**184**	**194**	**136**	**187**	**162**
K (Sample Size [# of Level 3 units]) in **PowerUpR**	**183**	**47**	**272**	**144**	**33**	**125**	**145**	**217**	**184**	**194**	**135**	**186**	**162**

Note. g_3 (number of covariates added at level 3) is fixed at 1 for all 12 designs.