The Journal of Experimental Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Statistical Power for Randomized Controlled Trials with Clusters of Varying Size

Joseph M. Kush, Timothy R. Konold & Catherine P. Bradshaw

To cite this article: Joseph M. Kush, Timothy R. Konold & Catherine P. Bradshaw (2022) Statistical Power for Randomized Controlled Trials with Clusters of Varying Size, The Journal of Experimental Education, 90:3, 673-692, DOI: 10.1080/00220973.2021.1873089

To link to this article: https://doi.org/10.1080/00220973.2021.1873089

Published online: 15 Feb 2021.

THE JOURNAL OF EXPERIMENTAL EDUCATION
2022, VOL. 90, NO. 3, 673–692
https://doi.org/10.1080/00220973.2021.1873089

Routledge
Taylor & Francis Group

MEASUREMENT, STATISTICS, AND RESEARCH DESIGN

Statistical Power for Randomized Controlled Trials With Clusters of Varying Size

Joseph M. Kush, Timothy R. Konold, and Catherine P. Bradshaw

University of Virginia, Charlottesville, VA, USA

ABSTRACT

In two-level designs, the total sample is a function of both the number of Level 2 clusters and the average number of Level 1 units per cluster. Traditional multilevel power calculations rely on either the arithmetic average or the harmonic mean when estimating the average number of Level 1 units across clusters of unbalanced size. The current study compares these two approaches with simulation-based power estimates in cluster randomized controlled trial designs with unbalanced cluster size. Results from the Monte Carlo study demonstrated that the largest differences in simulated and calculated power occurred in study designs with large variability in the number of Level 1 units sampled. We discuss implications of these findings for the design of cluster randomized trials.

KEYWORDS

Harmonic mean; multilevel models; power; sample size; unbalanced clusters

RANDOMIZED CONTROLLED TRIALS (RCTs) are recognized as the "gold standard" for assessing an intervention's effectiveness (Institute of Education Sciences, 2003; Shadish et al., 2011). In the simplest RCT design, individuals are randomly assigned to an intervention group or to a control group, with the goal of estimating the effects of an intervention. Group, or cluster, randomized trials (CRTs) are increasingly prevalent as a means of designing evaluations of treatments in which nested data structures are present (Murray, 1998). In these designs, the higher-level units (clusters) are randomly assigned to treatment or control conditions. In educational settings, it is common for CRTs to involve the nesting of students within classrooms or teachers within schools. In public health CRTs, it might involve individuals nested in communities or neighborhoods; whereas in health CRTs, patients might be nested within hospitals (Fitzmaurice et al., 2011). In CRTs, randomization equates entire clusters across treatment and control conditions on all pretreatment variables (Berk, 2005). As such, CRTs are inherently multilevel in nature.

CRTs are quickly becoming the norm in educational and public health research. For example, Atkinson and Wade (2015) evaluated the effects of a mindfulness-based intervention in the prevention of eating disorders by randomly assigning 19 high school classrooms to intervention or control. Another educational example involved 37 elementary schools that were randomly assigned to a control condition or to a schoolwide prevention strategy, Positive Behavioral Interventions and Supports, aimed at reducing disruptive behavior problems, to evaluate student and staff outcomes (see Bradshaw et al., 2008). Yet another example of a CRT involved 22 clusters of primary care sites that were randomly assigned to a control condition or to the Sustained Patient-Centered Alcohol-Related Care prevention program, which aimed to address unhealthy alcohol use (Glass et al., 2018).

CONTACT Joseph M. Kush jmk7cj@virginia.edu School of Education and Human Development, University of Virginia, 207 Ridley Hall Charlottesville, VA 22904.

In practice, CRTs often have an unbalanced cluster size due to sampling designs, variation in consent rates, or eligibility criteria. For example, schools may vary in the number of classrooms within the school building, thus, creating imbalance among Level 1 units across clusters. In other studies, eligibility criteria (e.g., children with special education needs, Autism, or an Individualized Education Program) may result in considerable variability in the Level 1 units across schools. It is also possible that low or variable consent rates may contribute to unbalanced Level 1 units across the clusters. These and other such issues may make it challenging to estimate power, particularly when the number of Level 1 units is small. To date, there has been limited consideration of the impact of small and variable cluster sizes on power within the context of CRTs. Having an enhanced understanding of the effects of these types of varying parameters on power calculations may prove useful both for designing CRT studies and for determining the power to detect a significant effect in trials after they are fielded and to experience these real-world implementation challenges.

Evaluating treatment effects in CRTs

Data arising from two-group, two-level CRTs can be evaluated through the following equations, expressed here in hierarchical form:

$$\begin{aligned} Level\ 1: &\ y_{ij} = \beta_{0j} + e_{ij},\ e_{ij} \sim N(0, \sigma^2) \\ Level\ 2: &\ \beta_{0j} = \gamma_{00} + \gamma_{01} \times T_j + u_{0j},\ u_{0j} \sim N(0, \tau), \end{aligned} \tag{1}$$

where, at Level 1, y_{ij} represents the observed outcome y for unit i in cluster j, β_{0j} is the mean outcome for cluster j, and e_{ij} is an error term for each Level 1 unit that is assumed to follow a normal distribution with a mean of zero and a within-cluster variance, σ^2. At Level 2, γ_{00} is the grand mean outcome, γ_{01} is the mean difference in the outcome between treatment and control clusters, T_j is a binary treatment indicator variable for cluster j, and u_{0j} is a random effect term for cluster j that is assumed to follow a normal distribution with a mean of zero and a between-cluster variance, τ.

In two-group, two-level CRTs, researchers are often interested in power calculations regarding the treatment effect, γ_{01} :

$$\hat{\gamma}_{01} = \bar{Y}_T - \bar{Y}_C, \tag{2}$$

where $\overline{Y}_T$ and $\overline{Y}_C$ represent mean outcome values for the treatment group and control group, respectively. When there is an equal probability for study participants to be assigned to treatment or control conditions, the variance of the treatment effect can be estimated as

$$Var(\hat{\gamma}_{01}) = \frac{4\left(\tau + \frac{\sigma^2}{n}\right)}{J}, \tag{3}$$

where n is the number of units per cluster and J is the number of clusters (Raudenbush, 1997). A nondirectional statistical hypothesis of

$$\begin{aligned} H_0: &\ \gamma_{01} = 0 \\ H_a: &\ \gamma_{01} \neq 0 \end{aligned}$$

can be evaluated through an F statistic that can be derived from a two-factor ANOVA model (Kirk, 1982):

$$F = \frac{MS_T}{MS_C}, \tag{4}$$

where MS_T represents the mean squares for the treatment groups (fixed factor) and MS_C represents the mean squares for the clusters (random factor). As the number of clusters J increases,

the F statistic converges to the following ratio of expected mean squares:

$$\frac{E(MS_T)}{E(MS_C)} = 1 + \lambda,$$

where

$$\lambda = \frac{nJ\gamma_{01}^2/4}{n\tau + \sigma^2}. \tag{5}$$

When the null hypothesis is false, the F statistic follows a noncentral F distribution with one degree of freedom in the numerator, J-2 degrees of freedom in the denominator, and a noncentrality parameter λ :

$$\lambda = \frac{\gamma_{01}^2}{4\left(\tau + \frac{\sigma^2}{n}\right)/J}. \tag{6}$$

In balanced designs, when the numbers of clusters are equal across experimental conditions, estimation of the two-level CRT model using restricted maximum likelihood (REML) matches results of the nested ANOVA. However, REML estimation is better suited for instances in which the number of clusters varies across conditions (Raudenbush, 1993).

To give more meaning to parameters, variability can be redefined in terms of the intraclass correlation coefficient, ρ :

$$\rho = \frac{\tau}{\tau + \sigma^2}. \tag{7}$$

Here τ is equal to the between-cluster variance, σ^2 is equal to the within-cluster variance, and $\tau + \sigma^2$ is equal to the total variance. The intraclass correlation coefficient can be interpreted as the proportion of variance in the outcome that is between clusters or, more generally, an indicator of the degree of clustering. Similarly, the treatment effect can be standardized, δ :

$$\delta = \frac{\gamma_{01}}{\sqrt{\tau + \sigma^2}}. \tag{8}$$

Here, γ_{01} is equal to the difference in population means between treatment and control groups. Thus, the estimated standardized effect size, $\hat{\delta}$, can be estimated by

$$\hat{\delta} = \frac{\bar{Y}_T - \bar{Y}_C}{\sqrt{\tau + \sigma^2}}, \tag{9}$$

where $\overline{y}_T$ and $\overline{y}_C$ represent the mean outcomes for the treatment and control groups, respectively. The standardized effect size can be interpreted as the standard deviation difference between mean outcomes for the treatment and control groups.

Power in CRTs

The probability of being able to reject the null hypothesis of no treatment effect when one exists is referred to as the power of a test. Given the significant time, effort, and costs associated with conducting CRTs, it is important that researchers design the trial to be adequately powered to detect a treatment effect for a particular design, estimated effect size, and projected sample size. Power in multilevel models is affected by multiple factors including the significance level α, the treatment effect δ, the intraclass correlation coefficient ρ, the number of Level 2 clusters, and the number of Level 1 units per cluster (n_i; Spybrook et al., 2011).

Among these, the number of clusters and the number of units per cluster are likely to be the most malleable of the factors affecting power that are in the researchers' control. For this reason,

the effects of sampling decisions on power within multilevel frameworks has remained an active area of research (e.g., Cox & Kelcey, 2019; Kelcey et al., 2019; Konstantopoulos, 2010; Usami, 2014). For example, research has demonstrated that increasing the number of J clusters sampled improves power more than increasing the number of i units per cluster (Snijders & Bosker, 1993; Spybrook et al., 2011). However, the sampling of additional clusters often comes at a greater financial cost than the sampling of more i units per cluster and could be less realistic in applied settings. See for example, Raudenbush and Liu (2000) and Snijders and Bosker (1999) for additional discussion of optimal costs and efficient allocations of resources in multilevel designs.

One area of methodological focus has been on how different numbers of clusters assigned to experimental conditions impact estimates of power. For instance, Konstantopoulos (2010) examined the effects on power in CRT designs when sample sizes between treatment and control groups differed. Results indicated that power estimates for unbalanced designs were smaller than those from balanced designs (i.e., equal number of clusters in treatment and control). Similarly, Liu (2003) considered unbalanced designs in terms of unequal sample allocation between treatment and control units and effects on costs per sampling unit. In the aggregate he found that statistical power may be higher for unbalanced designs as compared to balanced designs if the control condition costs significantly less money than the intervention condition. As such, optimal power occurs when approximately 75% of all clusters are assigned to control, with only 25% of clusters assigned to treatment.

The issue of unbalanced cluster size

As noted above, in multilevel CRTs, the total sample size is a function of both the number of Level 2 clusters and the number of Level 1 units within clusters. In most power calculations, estimates of the total sample size are typically obtained as the product of the number of Level 2 clusters (J) multiplied by the average number of Level 1 units per cluster ($\overline{n}_i$) where the arithmetic average number of Level 1 units per cluster, $\overline{n}_i$, is equal to the total number of Level 1 units sampled (i) divided by the total number of clusters sampled (J):

$$\overline{n}_i = \frac{\sum i}{J}. \tag{10}$$

Use of the arithmetic average number of Level 1 units, however, might be inappropriate for scenarios in which the number of Level 1 units varies widely across clusters. For example, consider two different CRT designs. In Study A, exactly 10 units are sampled from each of 50 clusters. In Study B, two units are selected from each of 49 clusters and 402 units are selected from one cluster. In both studies, a total of $J = 50$ clusters was selected, and both have the same average number of units per cluster, $\overline{n}_i = 10$. While there is no variability in the number of units per cluster sampled in Study A, there is extreme variability in the number of units per cluster in Study B ($SD = 56$). Yet traditional power calculations, which assume an average number of Level 2 units per cluster, would yield the same power estimates for both studies, holding all other factors constant. However, the standard errors in Study B would be materially influenced by the unbalanced nature of the single cluster in which 402 Level 1 units were samples. For this reason, researchers have suggested replacing the arithmetic average with the harmonic mean to more closely approximate the standard error of a treatment effect when cluster sizes are unequal (Cohen, 1988; Kelcey et al., 2019; Raudenbush, 1997; Spybrook et al., 2011), where the harmonic mean number of Level 1 units per cluster ($\overline{n}_{i_H}$) is equal to the total number of clusters (J) divided by the summation of the reciprocal of each cluster sample size (i) across all clusters:

$$\overline{n}_{i_H} = \frac{J}{\sum_{j=1}^{J} i^{-1}}. \tag{11}$$

In the example above, the $\overline{n}_{i_H} = 10$ in Study A, while $\overline{n}_{i_H} = 2.04$ for Study B. The harmonic mean reflects the meaningful differences in sampling designs between the two fictitious studies.

Several simulation-based studies of power in CRTs have focused on the performance of power estimates in the context of the arithmetic average of the projected cluster size (Maas & Hox, 2005; Scherbaum & Ferreter, 2009). For example, in school-based work, this is often done by estimating the average number of students or teachers per school that are likely to enroll in the project. However, simulation work by Manatunga et al. (2001) demonstrated that use of the arithmetic average in CRT power calculations underestimated the sample size required when dealing with large variations in Level 1 units across clusters. As a result, the researchers proposed a correction term in which the total number of clusters sampled increases as the variability in cluster sizes increases. Others recommend increasing the number of clusters sampled by 10% to correctly account for variation in cluster size (Van Breukelen et al., 2007). Given the cost and burden associated with conducting CRTs, these adjustments to the study design need to be carefully considered, and more precise estimates are needed to ensure adequate power in the context of real-world situations in which imbalance is likely to occur. As such, there is need for improved understanding of the impact of variation in the Level 1 units on power in CRTs.

Current study

Much of the research on power in the context of CRTs has focused on the effects on estimates derived from the arithmetic mean, the harmonic mean, and/or proposed correction terms. Yet little is known about how CRTs with variability in the number of Level 1 units per cluster compare with power calculations that are based on either the arithmetic average or the harmonic mean number of Level 1 units per cluster. The current study evaluated how variation in the number of Level 1 units per cluster impacts statistical power in the context of two-level CRT designs through the use of a Monte Carlo simulation study. Specifically, we sought to address the following research question: *How do traditional calculations of multilevel power utilizing the arithmetic average number of Level 1 units per cluster or the harmonic mean compare to simulation-based estimates of power for two-level cluster randomized trials with clusters of varying size?*

To address this research question, we used Monte Carlo simulations to generate data from a predefined population with fixed parameter values. Models were fit to the data and estimates of the treatment effect were used to determine the simulation-based estimates of power. These estimates of power were then compared to calculations of power that utilized the arithmetic average number of Level 1 units per cluster and the harmonic mean-based number of Level 1 units per cluster to understand more about the role of variability in cluster sample size and its relation to statistical power. We were particularly interested in the lower end of this range, such as in the context of CRTs with few Level 1 units (e.g., students with special education needs in a school or few early-career teachers within a school), as we anticipated that imbalance in these situations would lead to more-biased estimates of power than in studies with a large number of Level 1 units. We also considered the impact of variation of several parameters, such as the total number of Level 2 clusters, the intraclass correlation coefficient (ICC), and the effect size within the context of the Monte Carlo simulation.

Method

Simulation study

Multilevel data sets were simulated for two-level CRT designs, in which Level 1 units were nested within Level 2 clusters and treatment was assigned at the cluster level. Outcome values for Level 1 units were generated through the following equation:

$$Y_{ij} = \delta \times T_j + u_{0j} + e_{ij}. \tag{12}$$

Here, Y_{ij} represents an observed continuous outcome value for unit i in cluster j, δ represents a standardized effect size associated with being in a treatment cluster, T_j represents a binary treatment indicator variable for cluster j, u_{0j} represents a random effect term for cluster j, and e_{ij} represents an error term for each Level 1 unit. Importantly, u_{0j} followed a normal distribution with a mean equal to zero and a variance τ equal to a specified intraclass correlation coefficient ρ. Similarly, e_{ij} followed a normal distribution with a mean equal to zero and a variance equal to 1 ρ. Moreover, all clusters were assumed to have an equal probability of being assigned to treatment or control conditions, such that $P(T_j = 1) = 0.5$. Data sets were generated using R 2.3.1 software (R Core Team, 2020).

Following data generation, a mixed-effects model was fit to the data, in which the average treatment effect estimate was the parameter of interest. All models were estimated using REML estimation to appropriately handle unbalanced cluster sizes (Raudenbush, 1993). The power to detect a treatment effect is defined as the proportion of replications for which the null hypothesis, that the parameter is equal to zero, is rejected at a given significance level. We used the .05 significance level (two-tailed test with a critical value equal to 1.96) for all models. Models were fit using the lme4 package in R (Bates et al., 2015).

Power estimates obtained through the use of the arithmetic average number of Level 1 units per cluster and the harmonic mean number of Level 1 units per cluster were then compared to simulation-based estimates of power. As previously noted, the computation of power for the average treatment effect uses an F statistic that follows a noncentral F distribution with one degree of freedom in the numerator and J-2 degrees of freedom in the denominator, with the noncentrality parameter λ. Let F_{CV} represent the critical value of F for a nondirectional test with a significance level of .05. Then, power for the model presented in Equation 12 was calculated as:

$$Power = 1 - \beta,$$

where

$$\beta = Prob[F(1, J-2; \lambda) < F_{CV}\ ]. \tag{13}$$

Design facets

The Monte Carlo simulation contrasted a total of five design facets: (a) the number of Level 2 units J, (b) the ICC ρ, (c) the standardized effect size δ, (d) the minimum number of Level 1 units per cluster, and (e) the maximum number of Level 1 units per cluster. This resulted in a total of 1,260 unique simulation cells. All design facets are presented in Table 1.

Regarding the number of Level 2 clusters, prior research by Hox and Maas (2001) and Maas and Hox (2005) found cluster sizes of fewer than 50 may lead to biased estimates in multilevel structural equation models. For simpler observed multilevel models, recommendations for more than 10 clusters (Snijders & Bosker, 1993) and more than 30 clusters (Hoyle & Gottfredson, 2015) have been made to ensure reliable estimates. A recent review of 49 empirical studies examining school-level treatment effects for CRTs revealed a range of $J = 11$ to $J = 60$ sampled schools for two- and three-level designs (Spybrook, 2014). In these studies, the mean number of schools was 30.4 ($SD = 16.7$, median = mode = 30). Given that power approaches 1 as the number of clusters increases, regardless of other factors (Bloom, 2005; Spybrook et al., 2011), we anticipated that any differences between calculated power and simulation-based estimates of power would decrease toward 0 as J increased toward infinity. Based on these considerations, we fit models with four different sample sizes of Level 2 units: $J = 20$, 30, 50, and 60.

With regard to variation in the ICCs, research on educational interventions and meta-analyses have reported ICC values typically ranging from 0.05 to 0.25 (Bloom et al., 1999; 2007; Hedges &

Table 1. Summary of Monte Carlo population specifications for two-level CRT designs with variability in Level 1 units.

Level 2 Sample Size	$J = 20, 30, 50, 60$
ICC	$\rho = 0.05, 0.10, 0.20$
Effect Size	$\delta = 0.2, 0.3, 0.4$
Minimum Level 1 Sample Size	$n_i = 5, 10, 15, 20, 30$
Maximum Level 1 Sample Size	$n_i = 10, 15, 20, 25, 30, 40, 50$

Hedberg, 2007; Murray & Short, 1995). Therefore, we chose three different ICCs: $\rho = 0.05$, 0.10, and 0.20 within this range.

Regarding standardized effect sizes, three different values were selected: $\delta = 0.20$, 0.30, and 0.40. These values can be interpreted analogously to Cohen's d (Cohen, 1992), in which effects less than 0.20 are considered small, from 0.20 to 0.5 are considered moderate, and above 0.50 are considered large. In a meta-analysis of 61 school-level CRTs, Hill et al. (2007) found mean effect sizes in the 0.20 to 0.30 range. Similarly, Spybrook et al. (2016) conducted a meta-analysis of 38 school-level CRTs in which the minimum detectable effect size for the two cohorts considered is 0.48 and 0.23, respectively, ultimately suggesting that educational CRTs be designed to detect effect sizes in the range of 0.20 to 0.30.

Lastly, we varied both the minimum and maximum number of Level 1 units per cluster. Specifically, we considered five values for the minimum number of Level 1 units per cluster: $n_{min} = 5$, 10, 15, 20, and 30. We also considered seven values for the maximum number of Level 1 units per cluster: $n_{max} = 10$, 15, 20, 25, 30, 40, and 50. In all scenarios, the minimum and maximum number of Level 1 units per cluster were evenly split between clusters. For scenarios in which the minimum number of Level 1 units per cluster was equal to the maximum number, the arithmetic average number of Level 1 units per cluster was equal to the harmonic mean. In total, 5,000 iterations were conducted for each unique simulation condition.

Results

Simulation-based estimates of power were calculated as the proportion of iterations for which the null hypothesis, that the treatment effect is equal to 0, was rejected at the .05 significance level. The simulation-based estimates of power were then compared to calculations of power utilizing either the arithmetic average or harmonic mean number of Level 1 units per cluster. Values representing the relative difference between simulation-based estimates of power and calculated power are presented in Tables 2–5. To condense the amount of output provided in each table, while still providing key summaries, we restricted the maximum number of Level 1 units per cluster to 10, 25, and 50 in our tables for ease of presentation.

Several general findings emerged across all conditions and scenarios. The majority (77%) of power calculations based on the arithmetic average number of Level 1 units per cluster calculations were greater than the simulation-based estimates of power. This can be seen by the majority of positive values in the upper half of Tables 2–5, as this indicates that arithmetic average calculations of power tend to overestimate true power. By contrast, the majority (77%) of power calculations based on the harmonic mean number of Level 1 units per cluster were less than the simulation-based estimates of power. Similarly, this can be seen by the majority of negative values in the lower half of Tables 2–5, as this indicates that harmonic mean calculations of power tend to underestimate true power.

Differences between the arithmetic average and the harmonic mean number of Level 1 units per cluster were also related to differences in power. For example, the larger the difference between the arithmetic average and the harmonic average number of Level 1 units per cluster, the larger the difference in power between simulation-based estimates and calculations

Table 2. Differences between calculated and simulated power for two-level CRT designs for $J = 20$ clusters.

		Calculated Power Using Arithmetic Average Minus Simulated Power								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i/j*	Max. *i/j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	.007	.007	.011	.003	.005	.009	−.005	.004	.006
	25	.026	.034	.061	.008	.032	.046	.008	.010	.026
	50	.046	.103	.085	.029	.065	.092	.020	.031	.048
10	10	.008	.005	.004	.007	.002	.006	.003	.008	.012
	25	.011	.032	.023	.001	.003	.006	.007	.004	.012
	50	.025	.045	.050	.011	.034	.026	.010	.003	.021
15	10	.011	.008	.001	.004	.001	.010	.004	.003	.010
	25	.010	.007	.006	.003	.005	.014	.001	.001	.010
	50	.018	.035	.028	.003	.023	.017	.004	.003	.012
20	10	.004	.003	.011	.001	.002	.018	.012	.006	.014
	25	.001	.005	.003	.006	.002	.003	.010	−.004	.003
	50	.014	.013	.009	.001	.006	.014	.002	.001	.008
30	10	.014	.036	.030	.002	.008	.020	.003	.004	.003
	25	.003	.008	.003	.007	.006	.003	.001	.001	.000
	50	.003	.011	.013	.004	.003	.004	.008	.004	.020
		Calculated Power Using Harmonic Mean Minus Simulated Power								
5	10	.002	−.015	−.021	−.004	−.008	−.012	−.008	−.003	−.005
	25	.031	−.082	−.094	−.022	−.034	−.052	−.004	−.017	−.018
	50	.063	−.109	−.166	−.023	−.045	−.068	.001	−.010	−.020
10	10	.008	.005	.004	.007	−.002	−.006	.003	.008	.012
	25	.010	−.009	−.024	−.011	−.024	−.024	−.010	−.011	−.001
	50	.034	−.061	−.060	−.014	−.018	−.045	.002	−.014	−.008
15	10	.007	.000	−.010	−.006	−.004	.003	−.005	.001	.007
	25	.003	−.005	−.007	−.006	−.011	.005	−.002	−.003	−.014
	50	−.016	−.023	−.026	−.010	−.005	−.019	−.008	−.006	−.002
20	10	−.008	−.027	−.018	−.005	−.014	−.001	.010	.001	.006
	25	.000	.003	−.005	.006	−.003	−.005	−.010	−.004	.003
	50	−.005	−.019	−.019	−.006	−.009	−.005	−.004	−.003	.000
30	10	−.015	−.021	−.034	−.015	−.020	−.021	−.008	−.007	−.013
	25	−.004	−.009	.001	.006	.005	−.003	.001	−.001	−.001
	50	−.002	.002	.006	−.006	−.002	−.001	−.009	−.006	.018

Note. The maximum number of Level 1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

utilizing the arithmetic average ($r = -.73$, $p < .01$). Similarly, the larger the difference between the arithmetic average and the harmonic average number of Level 1 units per cluster, the larger the difference in power between simulation-based estimates and calculations utilizing the harmonic mean ($r = .60$, $p < .01$). Taken together, these findings suggest that calculations of power utilizing either the arithmetic average or harmonic mean number of Level 1 units per cluster may overestimate or underestimate, respectively, the true power of a model when cluster sizes are unbalanced.

With regard to variation in the number of Level 1 units, holding constant all factors other than Level 1 sample sizes, the largest difference in power between simulated power and arithmetic average power occurred in instances in which there were a minimum of five per cluster and a maximum of 50 units per cluster, representing large Level 1 variability. Here, simulated power was less than the arithmetic average calculation. For example, results for $J = 20$ clusters are shown in Table 2. The largest difference in power between simulated power and harmonic mean power also occurred with the minimum of five and maximum of 50 units per cluster scenario, where simulated power was greater than the harmonic mean calculation. For example, the largest difference in arithmetic average calculations of power and simulated power occur for the scenario with a minimum of five and maximum of 50 units per cluster, an ICC of 0.05, and an effect size of

0.3. By contrast, the smallest differences in power occurred for scenarios with almost no variation in Level 1 sample sizes for both the arithmetic average and harmonic mean calculations of power.

Similar results were found for $J = 30$ clusters, presented in Table 3. The largest differences in power, holding all else equal, occurred for the scenario with a minimum of five and maximum of 50 units per cluster, the scenario with the largest variability in Level 1 sample sizes. For example, the largest difference in harmonic mean calculations of power and simulated power occur for the scenario with a minimum of five and maximum of 50 units per cluster, an ICC of 0.05, and an effect size of 0.3. These differences in power were true for calculations utilizing either the arithmetic average or the harmonic mean. In general, simulated power in Table 3 was less than the calculation of power with the arithmetic average but greater than the calculation of power with the harmonic mean.

Next, we examined the models for $J = 50$ clusters (see Table 4). Again, holding all facets constant, we see that the largest difference in power between simulated power and calculation of power using the arithmetic average occurred for the scenario with a minimum of five and maximum of 50 units per cluster, where simulated power is less than the calculation. The largest difference in power between simulated power and calculation of power using the harmonic mean again occurs for the same scenario with large variability in Level 1 units, where simulated power was greater than the calculation of power. More specifically, the largest differences, as shown in Table 4, can be seen for the scenario with a minimum of five and maximum of 50 units per

Table 3. Differences between calculated and simulated power for two-level CRT designs for $J = 30$ clusters.

		Calculated Power Using Arithmetic Average Minus Simulated Power								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i/j*	Max. *i/j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	.004	.020	.010	.007	.001	.012	.000	.015	.013
	25	.031	.050	.059	.023	.045	.059	.005	.031	.037
	50	.065	.100	.067	.052	.086	.093	.016	.041	.068
10	10	.010	.005	.002	−.007	.014	.008	−.011	.001	−.009
	25	.014	.026	.017	.001	.020	.027	.009	.007	−.002
	50	.045	.056	.030	.025	.027	.035	.010	.020	.028
15	10	−.001	.005	−.004	−.002	−.002	.005	−.002	.003	.005
	25	−.002	−.007	.010	.007	.012	.013	−.002	.004	.001
	50	.029	.030	.019	.003	.008	.021	−.004	.010	.006
20	10	−.004	.011	.010	.005	.025	.009	.012	−.002	.005
	25	−.002	.000	.002	−.003	.020	−.001	−.007	.004	.002
	50	.018	.021	.005	.016	.020	.001	−.002	.003	.010
30	10	.018	.035	.017	.008	.019	.039	.000	.009	.022
	25	.021	.008	.001	−.003	−.012	.000	.004	−.006	.005
	50	.005	.005	.002	.002	.006	.009	−.006	−.005	.005
		Calculated Power Using Harmonic Mean Minus Simulated Power								
5	10	−.019	−.010	−.027	−.003	−.019	−.016	−.004	.005	−.003
	25	−.057	−.102	−.084	−.025	−.048	−.054	−.014	−.010	−.024
	50	−.101	−.153	−.130	−.030	−.066	−.076	−.014	−.022	−.024
10	10	.010	.005	.002	−.007	.014	.008	−.011	.001	−.009
	25	−.018	−.022	−.018	−.015	−.009	−.004	.003	−.004	−.019
	50	−.042	−.058	−.037	−.014	−.042	−.032	−.002	−.007	−.009
15	10	−.008	−.006	−.013	−.005	−.009	−.003	−.003	.000	.001
	25	−.012	−.021	.002	.002	.003	.004	−.003	.000	−.004
	50	−.020	−.027	−.010	−.018	−.027	−.012	−.010	−.003	−.012
20	10	−.023	−.018	−.013	−.004	.007	−.011	.008	−.009	−.006
	25	−.004	−.003	.001	−.004	.018	−.003	−.007	.004	.001
	50	−.009	−.009	−.008	.004	.001	−.016	−.006	−.004	.000
30	10	−.027	−.030	−.028	−.014	−.020	−.002	−.007	−.007	.000
	25	.019	.006	.000	−.003	−.013	−.001	.004	−.006	.005
	50	−.003	−.003	−.001	−.001	.001	.004	−.007	−.006	−.007

Note. The maximum number of Level 1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

Table 4. Differences between calculated and simulated power for two-level CRT designs for $J = 50$ clusters.

		Calculated Power Using Arithmetic Average Minus Simulated Power								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i/j*	Max. *i/j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	.004	.009	.006	.012	.010	.005	.000	.017	−.005
	25	.055	.046	.013	.037	.055	.041	.018	.045	.039
	50	.103	.073	.013	.065	.085	.051	.031	.058	.069
10	10	.000	.000	−.002	.013	−.008	−.003	.004	.003	.006
	25	.017	.018	.004	.022	.019	.010	−.001	.016	.010
	50	.055	.024	.002	.037	.044	.017	.025	.032	.033
15	10	.012	−.001	.001	−.002	.000	.009	−.004	−.007	.012
	25	−.002	−.005	.001	−.004	.013	.003	.004	.006	−.005
	50	.041	.017	.002	.036	.019	.008	.006	.025	.019
20	10	−.002	.013	.005	.016	.004	.004	−.002	.001	.007
	25	−.002	−.002	.000	.006	.009	.001	.002	.003	−.008
	50	.009	.003	.002	.002	.014	.005	−.002	.013	.016
30	10	.022	.016	.005	.039	.023	.013	.005	.008	.009
	25	−.006	.003	−.002	−.013	.000	−.003	−.001	.005	−.007
	50	.002	.003	.000	.002	.017	.001	−.001	−.004	.000
		Calculated Power Using Harmonic Mean Minus Simulated Power								
5	10	−.022	−.028	−.020	−.004	−.017	−.019	−.008	.001	−.024
	25	−.080	−.099	−.048	−.041	−.059	−.036	−.014	−.016	−.027
	50	−.135	−.128	−.050	−.066	−.086	−.049	−.020	−.033	−.024
10	10	.000	.000	−.002	.013	−.008	−.003	.004	.003	.006
	25	−.028	−.018	−.005	−.002	−.012	−.006	−.010	−.001	−.007
	50	−.060	−.045	−.009	−.023	−.025	−.014	.003	−.005	−.003
15	10	.003	−.010	−.002	−.008	−.008	.005	−.007	−.011	.008
	25	−.015	−.014	−.001	−.011	.005	−.001	.001	.001	−.010
	50	−.020	−.013	−.001	.004	−.015	−.005	−.005	.007	.002
20	10	−.029	−.010	−.002	.001	−.016	−.007	−.008	−.010	−.004
	25	−.005	−.004	−.001	.005	.007	.001	.002	.002	−.008
	50	−.024	−.011	.001	−.015	−.003	−.001	−.007	.004	.007
30	10	−.041	−.030	−.005	.005	−.018	−.008	−.008	−.014	−.013
	25	−.008	.003	−.002	−.014	.000	−.003	−.001	.004	−.007
	50	−.007	−.001	.000	−.003	.013	−.001	−.002	−.007	−.003

Note. The maximum number of level-1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

cluster, an ICC of 0.05, and an effect size of 0.2. The smallest differences in power for both simulated power versus arithmetic-average-calculated power and simulated power versus harmonic-mean-calculated power were observed in scenarios with small variability in cluster size.

Finally, results for $J = 60$ clusters are reported in Table 5. Holding all other factors equal, the largest differences in power occur for the scenario with a minimum of five and maximum of 50 units per cluster, representing the scenario with the largest variability in Level 1 sample sizes. While simulated power was less than calculated power using the arithmetic average, simulated power was greater than the calculated power using the harmonic mean. The smallest differences in power again occur for scenarios with very little variability in Level 1 units. For example, the difference in harmonic mean calculated power and simulated power is less than 0.001 for the scenario with a minimum of 20 and maximum of 25 units per cluster, an ICC of 0.1, and an effect size of 0.4.

It is important to note that overall, any differences between simulation-based estimates of power and calculations of power using the arithmetic average were not systematically related to Level 2 sample size ($r = .02$, $p = .58$) or effect size ($r = .02$, $p = .52$) but were related to the ICC ($r = -.14$, $p < .01$), although the relative magnitude of the correlation appears small. Similarly, differences between simulation-based estimates of power and calculations of power using the harmonic mean were not systematically related to Level 2 sample size ($r = .01$, $p = .68$) or effect size ($r = .01$, $p = .98$) but were related to the ICC ($r = .27$, $p < .01$), again relatively small in magnitude. However, the difference in the number of Level 1 units per cluster

Table 5. Differences between calculated and simulated power for two-level CRT designs for $J = 60$ clusters.

		Calculated Power Using Arithmetic Average Minus Simulated Power								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i/j*	Max. *i/j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	.008	.009	.012	.001	.015	.011	.005	.014	.005
	25	.053	.038	.008	.056	.064	.024	.020	.030	.036
	50	.100	.040	.004	.083	.086	.032	.046	.077	.055
10	10	−.009	.001	−.003	.004	.004	−.005	.003	−.007	.007
	25	.015	.014	.002	.013	.010	.006	.012	.019	.005
	50	.058	.015	.001	.057	.036	.012	.018	.029	.011
15	10	.014	.005	.000	.001	.009	.002	−.008	.005	.006
	25	.011	.005	.000	.011	.007	.003	−.001	.016	−.002
	50	.023	.007	.000	.024	.019	.002	.010	.015	.008
20	10	.016	.009	.001	−.003	.006	.003	.006	.011	.003
	25	.007	.000	−.001	−.002	.002	.001	−.004	.002	.008
	50	.023	.006	.000	.009	.019	.000	.009	.008	.003
30	10	.020	.017	.002	.029	.031	.008	.018	.018	.013
	25	.004	−.003	.000	−.005	.002	.003	.004	−.002	−.001
	50	.006	.002	.000	.002	−.003	.000	.001	.014	−.010
		Calculated Power Using Harmonic Mean Minus Simulated Power								
5	10	−.021	−.027	−.005	−.019	−.014	−.009	−.004	−.003	−.012
	25	−.096	−.084	−.026	−.035	−.045	−.030	−.019	−.036	−.023
	50	−.153	−.115	−.027	−.066	−.070	−.033	−.014	−.020	−.026
10	10	−.009	.001	−.003	.004	.004	−.005	.003	−.007	.007
	25	−.033	−.012	−.001	−.015	−.018	−.004	.001	.001	−.009
	50	−.058	−.030	−.002	−.010	−.023	−.005	−.007	−.010	−.018
15	10	.004	−.003	−.002	−.006	.002	−.001	−.011	.001	.002
	25	−.002	−.001	−.001	.003	.000	.001	−.004	.011	−.005
	50	−.036	−.010	−.001	−.011	−.009	−.005	−.003	−.004	−.006
20	10	−.013	−.009	−.002	−.020	−.012	−.004	−.001	−.001	−.006
	25	.004	−.001	−.001	−.003	.001	.000	−.005	.001	.007
	50	−.009	−.002	.000	−.010	.005	−.003	.003	−.002	−.004
30	10	−.045	−.016	−.002	−.009	−.006	−.004	.003	−.005	−.005
	25	.003	−.004	.000	−.005	.001	.003	.004	−.002	−.002
	50	−.002	.000	.000	−.004	−.006	−.001	−.001	.012	−.012

Note. The maximum number of Level 1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

(i.e., $n_{max} - n_{min}$) was significantly related to differences in simulation-based estimates of power for both arithmetic average calculations of power ($r = .53$, $p < .01$) and harmonic mean calculations of power ($r = -.44$, $p < .01$). Taken together, the findings consistently suggested that the variability in Level 1 sample sizes was the primary source behind any differences in power between simulation-based estimates and calculations of power.

Discussion

CRTs have been the focus of considerable methodological work (Dong et al., 2018; Kelcey et al., 2019; Konstantopoulos, 2012; Schochet, 2008; Usami, 2014). The current study adds to the literature by considering designs with unequal Level 1 sample sizes across clusters. Previous work in this area suggests that use of the harmonic mean is preferred over the arithmetic mean when numbers of Level 1 units are not equal across clusters in CRT conducted in real-world settings such as schools. The current interrogation of this issue leveraged simulation-based estimates of power, thereby, enabling us to contrast power estimates that would be obtained through use of the arithmetic average or the harmonic mean across a range of parameters (e.g., ICC, Level 1 sample sizes). In the real world, clusters vary in size; this is almost always the rule rather than the exception. This may be due in part to a limited eligible participant pool, dropout or attrition of participants, and other recruitment issues (Groves et al., 2009; Manatunga et al., 2001). For

example, Resnicow et al. (1998) conducted a study in which 32 schools were randomized to a health promotion educational intervention or control. Here, all third-grade students in the same school received the same intervention or control condition. However, the number of third-grade students in each school ranged from 20 to 81. Other examples include Tolan et al. (2020) who sampled 188 teachers from 72 schools, with a range of 1–13 teachers sampled from each school, and Lam et al. (2015) who sampled 3,288 students in from 188 classrooms, with a range of 6–28 students sampled from each school.

As such, this paper advances prior work on statistical power for CRTs and multilevel designs more generally by examining how different estimates of cluster size can impact traditional power calculations when the number of units per cluster is not constant. To address these gaps in the research, we examined simulation-based estimates of statistical power for two-level CRTs with variability in Level 1 sample size. Our results suggest that differences between simulation-based estimates and calculation-based estimates of power increase as variability in Level 1 sample size increases. Holding all simulation study facets constant (i.e., ICC, effect size), the largest differences in power values occurred in instances in which there was a minimum of five units per cluster and a maximum of 50 units per cluster; more specifically, these power differences were most pronounced when the number of Level 1 units was small (i.e., below 20).

In such scenarios, calculated power utilizing the arithmetic average tended to overestimate true power, while calculated power utilizing the harmonic mean tended to underestimate true power. By contrast, the smallest differences in power values, holding all other facets constant, occurred for scenarios with the smallest variability in Level 1 sample sizes. We were particularly interested in the low end of the Level 1 sample size range to determine the point at which researchers should become concerned about such variability in the cluster sizes. Results demonstrated that calculations of power increasingly diverged from the true simulation-based power as the minimum number of Level 1 units per cluster decreased. Moreover, the larger the variability in Level 1 sample sizes, the larger the differences in power.

Limitations

There are a few general limitations to our study that are important to note. First, the Monte Carlo simulation setup for this study examined a variety of facets with plausible values determined from prior meta-analyses in educational research. However, we were not able to explore all possible variants for each facet. For instance, we considered variability in the number of Level 1 units per cluster toward the lower end of sample sizes. We did not consider a scenario with a minimum of $n_{min}= 50$ Level 1 units per cluster and a maximum of $n_{max}= 95$, as the impact of cluster size variability on power estimates likely diminishes after Level 1 sample size reaches a certain threshold. Similar arguments hold for the total number of clusters, as power has been shown to increase toward 1 as the total number of clusters increases, regardless of other factors (Bloom, 2005; Spybrook et al., 2011). While the current study explored scenarios with 20 clusters, Spybrook's (2014) review of empirical studies revealed school-level CRTs with fewer than 20 clusters. We therefore encourage future researchers to examine power for multilevel designs with fewer than 20 clusters.

Other decisions made in the design of this simulation study may potentially limit our findings. For example, we assumed an equal probability of being assigned to treatment or control, such that $P(T_j = 1) = 0.5$. This allowed for a balanced design in that the number of clusters in the treatment group was equal to the number of clusters in the control group. However, CRTs and randomized control trials more generally may not always have such balance in practice (see Liu, 2003). Thus, findings from this study may or may not replicate in future studies examining variability in Level 1 sample sizes for unbalanced designs with treatment and control groups differing on the number of clusters (or units) per group. However, the issue of variable cluster sizes on

power in CRTs has been understudied in the literature. As a result, there is a need for future research to examine scenarios in which the number of Level 1 units per cluster is not evenly split between groups.

Lastly, our study examined effects of unbalanced cluster sizes on power for a relatively simple two-level, two-group CRT design, in which a single binary treatment indicator was used to estimate the average treatment effect. This model could be extended, for example, by including covariates such as pretest scores at Level 1, while simultaneously accounting for information at Level 2. The inclusion of covariates is a common way to increase the precision and power of a study, and empirical work has demonstrated that the inclusion of Level 1 or Level 2 covariates produces similar improvements in power (Bloom et al., 2007; Spybrook et al., 2011). The improvement in power from the addition of a covariate would likely reduce the required sample size necessary for a desired level of power, where a reduction in Level 1 sample size would not reduce power as much as a reduction in Level 2 sample size (Snijders & Bosker, 1993; Spybrook et al., 2011). Another natural extension to the models considered in this study would be three-level models (see Dong et al., 2018). While we explored the effects of variability in Level 1 sample sizes and power estimates in a simplistic two-level CRT design, future work should also explore such variability for more nuanced modeling approaches and study designs.

Conclusions and implications

These findings highlight the need to carefully consider the impact of variation in the number of observations at Level 1 when designing CRTs. While different power software and online tools currently exist for calculating power in multilevel models (e.g., optimal design; Spybrook et al., 2011), these calculations rely on either the arithmetic average or the harmonic mean number of units per cluster. Simulation-based estimates of power may offer more flexibility for researchers designing and planning CRTs than power software and tools utilizing arithmetic average or harmonic mean calculations. Although we are hesitant to offer a rule of thumb as to when variation in Level 1 sample sizes becomes large enough to significantly differ from traditional power calculations, one substantively important framing of this issue involves comparing differences in cluster sample sizes required for differences in power calculations and simulated power. To explore this, additional power calculations were computed to help answer the question, How many more (or fewer) clusters are needed to sample for the original calculation of power to match the simulation-based estimate of power? Data representing the difference in number of clusters required to equal simulated power are presented in Tables A1–A4 in the Appendix.

Across all tables, the majority (71%) of values for the arithmetic calculations are equal to or less than zero, indicating that fewer clusters are needed to be sampled for the arithmetic calculation of power to equal the simulation-based estimate of power. For example, Table A2 shows a scenario with a minimum of five and maximum of 50 units per cluster, an ICC of 0.1, and an effect size of 0.2 in which it can be seen that six fewer clusters are needed for the original arithmetic average calculation of power to equal the true simulated power. In this scenario, researchers entering the average number of units per cluster into a power software calculator would conclude a value of power greater than the true power. As such, researchers failing to recognize this would have studies underpowered for what they believed to be the actual power for their studies. By contrast, the majority (79%) of values for the harmonic calculation were greater than zero, suggesting that the harmonic calculation underestimates true power and that more clusters would need to be sampled to equal the simulation-based estimate. Again, as an example, the scenario in Table A3 with a minimum of 10 and maximum of 50 units per cluster, an ICC of 0.05, and an effect size of 0.3 indicates that an additional seven clusters would need to be sampled for the original harmonic mean calculation of power to equal the true power.

Along with findings shown in Tables 2–5, the largest differences in the number of clusters required occurs for a minimum of five and maximum of 50 units per cluster, the scenario with the largest ratio in cluster size. Differences in cluster sample sizes required for power calculations to match simulation-based estimates of power were slightly related to original Level 2 sample size ($r = -.11$), effect size ($r = -.18$), and the intraclass correlation coefficients ($r = -.15$). However, differences in the number of Level 1 units per cluster (i.e., $n_{max} - n_{min}$) were strongly related to differences in required clusters ($r = .58$). This perspective provides an alternative lens for understanding that variability in cluster sizes is most strongly associated with differences in the number of clusters required for calculations of power to match those obtained from simulation-based estimates of power.

Issues related to sampling costs and increased burdens of data collection are directly related to differences cluster sample sizes. As variability in cluster size increases, researchers must be cognizant of the implications of differences in power estimates that can arise across different methods for capturing cluster-size differences and the number of Level 2 clusters needed to obtain a desired level of power. As described elsewhere (Liu, 2003), these aspects of planning CRT designs will have direct impacts on study costs. Cost calculations would vary depending on the researcher's choice of using the arithmetic average (fewer clusters needed) or harmonic mean (more clusters needed) to achieve a desired level of power. By contrast, simulation-based estimates of power may provide more precision for researchers examining issues of costs and power for CRTs as well as other multilevel designs.

Taken together, these findings highlight the need for researchers to explicitly investigate characteristics unique to their study design, particularly those related to small and unbalanced cluster sizes, in light of their potential impact on statistical power. Calculations of power for CRT designs may over- or underestimate the true power of a model, as such analytic approaches are often restricted in critical ways. Simulation-based approaches may provide a more nuanced understanding of the impact of these design facets on statistical power within the context of CRTs.

Acknowledgements

The authors thank members of our larger research team including Drs. Elise Pas, Rashelle Musci, and Ji Hoon Ryoo.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H150027 and R305A150221 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID

Joseph M. Kush http://orcid.org/0000-0003-0183-494X
Timothy R. Konold http://orcid.org/0000-0003-0092-9234
Catherine P. Bradshaw http://orcid.org/0000-0003-2048-3225

References

Atkinson, M. J., & Wade, T. D. (2015). Mindfulness-based prevention for eating disorders: A school-based cluster randomized controlled study. *International Journal of Eating Disorders*, *48*(7), 1024–1037. https://doi.org/10.1002/eat.22416

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), v067i01. https://doi.org/10.18637/jss.v067.i01

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology*, *1*(4), 417–433. https://doi.org/http://dx.doi.org/10.1007/s11292-005-3538-2

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russel Sage Foundation.

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. https://doi.org/10.1177/0193841X9902300405

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Bradshaw, C. P., Koth, C. W., Bevans, K. B., Ialongo, N., & Leaf, P. J. (2008). The impact of school-wide positive behavioral interventions and supports (PBIS) on the organizational health of elementary schools. *School Psychology Quarterly*, *23*(4), 462–473. https://doi.org/10.1037/a0012883

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cox, K., & Kelcey, B. (2019). Optimal design of cluster- and multisite-randomized studies using fallible outcome measures. *Evaluation Review*, *43*(3–4), 189–225. https://doi.org/10.1177/0193841X19870878

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

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal data analysis* (2nd ed.). John Wiley & Sons, Inc.

Glass, J. E., Bobb, J. F., Lee, A. K., Richards, J. E., Lapham, G. T., Ludman, E., Achtmeyer, C., Caldeiro, R. M., Parrish, R., Williams, E. C., Lozano, P., & Bradley, K. A. (2018). Study protocol: A cluster-randomized trial implementing sustained patient-centered alcohol- related care (SPARC trial). *Implementation Science*, *13*(1), 108. https://doi.org/10.1186/s13012-018-0795-9

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). John Wiley & Sons, Inc.

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

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*. MDRC. https://www.mdrc.org/sites/default/files/full_84.pdf

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hoyle, R., & Gottfredson, N. C. (2015). Sample size considerations in prevention research applications of multilevel modeling and structural equation modeling. *Prevention Science: The Official Journal of the Society for Prevention Research*, *16*(7), 987–996. https://doi.org/10.1007/s11121-014-0489-8

Institute of Education Sciences. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Coalition for Evidence-Based Policy. https://ies.ed.gov/ncee/pdf/evidence_based.pdf

Kelcey, B., Spybrook, J., & Dong, N. (2019). Sample size planning for cluster-randomized interventions probing multilevel mediation. *Prevention Science: The Official Journal of the Society for Prevention Research*, *20*(3), 407–418. https://doi.org/10.1007/s11121-018-0921-6

Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Brooks/Cole.

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education*, *78*(3), 291–317. https://doi.org/10.1080/00220970903292876

Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, *47*(3), 392–420. https://doi.org/10.1080/00273171.2012.673898

Lam, A. C., Ruzek, E. A., Schenke, K., Conley, A. M., & Karabenick, S. A. (2015). Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology*, *107*(4), 1102–1115. https://doi.org/http://dx.doi.org/10.1037/edu0000028

Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics*, *28*(3), 231–248. https://doi.org/10.3102/10769986028003231

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Manatunga, A. K., Hudgens, M. G., & Chen, S. (2001). Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*, *43*(1), 75–86. https://doi.org/10.1002/1521-4036(200102)43:1<75::AID-BIMJ75>3.0.CO;2-N

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press, Inc.
Murray, D. M., & Short, B. (1995). Intra-class correlation among measures related to alcohol use by young adults: Estimates, correlates, and applications in intervention studies. *Journal of Studies on Alcohol*, *56*(6), 681–692. https://doi.org/10.15288/jsa.1995.56.681
R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–495). Marcel Dekker.
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., & 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
Resnicow, K., Davis, M., Smith, M., Baranowski, T., Lin, L. S., Baranowski, J., Doyle, C., & Wang, D. T. (1998). Results of the TeachWell worksite wellness program. *American Journal of Public Health*, *88*(2), 250–257. https://doi.org/10.2105/AJPH.88.2.250
Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*(2), 347–367. https://doi.org/http://dx.doi.org/10.1177/1094428107308906
Schochet, P. Z. (2008). Statistical power for randomized assignment evaluation of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87. https://doi.org/10.3102/1076998607302714
Shadish, W. R., Rodolfo, G., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, *16*(2), 179–191. https://doi.org/10.1037/a0023345
Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*(3), 237–259. https://doi.org/10.3102/10769986018003237
Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE.
Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education*, *82*(3), 334–357. https://doi.org/10.1080/00220973.2013.813364
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 Educational Sciences. *International Journal of Research & Method in Education*, *39*(3), 255–267.
Spybrook, J., Bloom, H., Cogdon, R., Hill, C., Martinez, A., Raudenbush, S. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software*. http://hlmsoft.net/od/od-manual-20111016-v300.pdf
Tolan, P., Elreda, L. M., Bradshaw, C. P., Downer, J. T., & Ialongo, N. (2020). Randomized trial testing the integration of the Good Behavior Game and MyTeachingPartner™: The moderating role of distress among new teachers on student outcomes. *Journal of School Psychology*, *78*, 75–95. https://doi.org/10.1016/j.jsp.2019.12.002
Usami, S. (2014). Generalized sample size determination formulas for experimental research with hierarchical data. *Behavior Research Methods*, *46*(2), 346–356. https://doi.org/10.3758/s13428-013-0387-1
Van Breukelen, G. J. P., Candel, M. J. J., & Berger, M. P. F. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicenter trials. *Statistics in Medicine*, *26*(13), 2589–2603. https://doi.org/10.1002/sim.2740

Appendix A

Table A1. Differences in required clusters sampled for calculated power to equal simulated power with $J = 20$ clusters.

		Calculation of Power Using Arithmetic Average								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i*/*j*	Max. *i*/*j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	0	0	−1	1	0	−1	2	0	0
	25	−2	−2	−3	0	−2	−3	−1	−1	−2
	50	−3	−5	−5	−3	−4	−4	−3	−2	−3
10	10	0	0	−1	0	0	0	1	0	−1
	25	−1	−2	−2	1	0	−1	2	1	−1
	50	−2	−3	−3	−1	−2	−2	−1	0	−2
15	10	0	−1	−1	1	0	−1	2	0	−1
	25	0	−1	−1	1	0	−1	1	1	0
	50	−1	−2	−2	0	−1	−2	2	0	−1
20	10	0	0	−1	1	0	−1	−1	0	−1
	25	0	−1	−1	0	0	0	3	1	0
	50	−1	−1	−2	0	−1	−1	1	0	−1
30	10	−1	−2	−2	1	−1	−2	2	0	0
	25	0	0	−1	0	0	−1	1	1	0
	50	0	−1	−2	1	0	−1	2	1	−1
		Calculation of Power Using Harmonic Mean								
5	10	1	1	0	2	1	0	3	1	0
	25	4	5	3	4	2	2	2	2	1
	50	8	6	6	4	3	2	1	1	1
10	10	0	0	−1	0	0	0	1	0	−1
	25	1	0	0	2	1	0	3	1	0
	50	3	2	1	2	1	1	0	1	0
15	10	0	0	0	2	0	−1	2	0	0
	25	0	0	−1	1	1	−1	2	1	1
	50	1	0	0	2	0	0	2	1	0
20	10	1	1	0	1	1	0	−1	0	0
	25	0	−1	−1	0	0	0	3	1	0
	50	0	0	0	1	0	−1	2	1	0
30	10	2	0	0	3	1	0	3	1	0
	25	0	0	−1	0	0	−1	1	1	0
	50	0	−1	−1	1	0	−1	3	1	−1

Note. The maximum number of Level 1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

Table A2. Differences in required clusters sampled for calculated power to equal simulated power with $J = 30$ clusters.

		Calculation of Power Using Arithmetic Average								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i*/*j*	Max. *i*/*j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	1	−2	−1	0	0	−1	1	−1	−1
	25	−3	−4	−5	−3	−4	−5	−1	−3	−3
	50	−5	−7	−8	−6	−6	−7	−3	−4	−5
10	10	−1	−1	−1	2	−1	−1	3	0	0
	25	−1	−2	−3	0	−2	−3	−1	−1	−1
	50	−4	−5	−5	−3	−3	−4	−1	−2	−3
15	10	0	−1	−1	1	0	−1	1	0	−1
	25	0	0	−2	−1	−1	−2	1	0	−1
	50	−3	−3	−4	0	−1	−3	1	−1	−1
20	10	0	−1	−2	0	−2	−1	−1	0	−1
	25	0	−1	−1	1	−2	−1	2	0	−1
	50	−2	−2	−2	−2	−2	−1	1	0	−1
30	10	−2	−3	−3	−1	−2	−3	1	−1	−2
	25	−2	−1	−1	1	0	−1	0	0	−1
	50	−1	−1	−1	0	−1	−2	2	0	0
		Calculation of Power Using Harmonic Mean								
5	10	3	0	1	1	2	0	2	0	0
	25	7	7	5	4	3	2	4	1	1
	50	12	10	9	5	4	3	3	2	1
10	10	−1	−1	−1	2	−1	−1	3	0	0
	25	2	0	0	2	0	−1	0	0	0
	50	3	2	2	2	2	1	1	0	0
15	10	1	0	0	1	0	−1	1	0	−1
	25	1	0	−1	0	−1	−1	1	0	0
	50	1	1	0	2	1	0	2	0	0
20	10	2	0	0	1	−1	0	−1	1	0
	25	0	−1	−1	1	−2	−1	2	0	−1
	50	0	0	0	0	−1	0	2	0	−1
30	10	2	1	1	2	1	−1	2	0	−1
	25	−2	−1	−1	1	0	−1	0	0	−1
	50	0	−1	−1	0	−1	−1	2	0	0

Note. The maximum number of Level 1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

Table A3. Differences in required clusters sampled for calculated power to equal simulated power with $J = 50$ clusters.

		Calculation of Power Using Arithmetic Average								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i/j*	Max. *i/j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	−1	−2	−2	−2	−2	−2	0	−2	0
	25	−7	−7	−6	−6	−7	−8	−4	−6	−5
	50	−11	−13	−12	−9	−10	−12	−6	−7	−9
10	10	0	−1	0	−2	0	−1	0	−1	−1
	25	−3	−4	−3	−3	−3	−3	0	−2	−2
	50	−7	−6	−4	−5	−6	−6	−5	−4	−5
15	10	−2	−1	−1	0	−1	−3	1	0	−2
	25	−1	0	−1	0	−2	−2	−1	−1	0
	50	−5	−5	−5	−5	−3	−3	−1	−4	−3
20	10	0	−3	−3	−3	−1	−2	1	−1	−2
	25	0	0	1	−1	−2	−1	0	−1	0
	50	−2	−2	−5	−1	−3	−3	0	−2	−3
30	10	−3	−4	−4	−5	−4	−4	−1	−2	−2
	25	0	−2	6	1	−1	0	0	−1	0
	50	−1	−2	3	−1	−3	−1	0	0	−1
		Calculation of Power Using Harmonic Mean								
5	10	3	2	2	1	1	1	2	−1	2
	25	11	11	12	7	6	4	3	1	2
	50	19	17	17	11	9	6	4	3	1
10	10	0	−1	0	−2	0	−1	0	−1	−1
	25	3	1	2	0	0	0	2	−1	0
	50	6	7	9	2	1	2	−1	0	−1
15	10	−1	0	0	1	0	−2	2	1	−2
	25	1	1	1	1	−1	−1	0	−1	0
	50	1	1	1	−1	0	0	1	−2	−1
20	10	3	0	0	0	1	0	2	0	−1
	25	0	0	1	−1	−2	−1	0	−1	0
	50	2	2	−2	1	−1	0	1	−1	−2
30	10	4	3	3	−1	1	1	2	1	0
	25	0	−1	6	1	−1	0	0	−1	0
	50	0	−1	4	0	−3	0	0	0	−1

Note. The maximum number of Level 1 units per cluster has been"tx restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.

Table A4. Differences in required clusters sampled for calculated power to equal simulated power with $J = 60$ clusters.

		Calculation of Power Using Arithmetic Average								
		$\rho = 0.05$			$\rho = 0.10$			$\rho = 0.20$		
Min. *i*/*j*	Max. *i*/*j*	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
5	10	−2	−2	−4	0	−3	−3	−1	−2	−2
	25	−8	−9	−8	−9	−10	−9	−4	−5	−7
	50	−13	−13	−13	−12	−13	−14	−9	−11	−10
10	10	1	−1	1	−1	−1	0	−1	0	−2
	25	−3	−4	−4	−2	−3	−3	−3	−3	−2
	50	−8	−7	−8	−8	−7	−8	−4	−5	−3
15	10	−3	−2	0	0	−2	−2	2	−1	−2
	25	−2	−2	2	−2	−2	−2	0	−3	−1
	50	−4	−5	5	−4	−4	−2	−2	−3	−3
20	10	−3	−3	−2	0	−2	−2	−1	−2	−2
	25	−2	−1	5	0	−1	−1	1	−1	−2
	50	−4	−4	−1	−2	−5	−1	−2	−2	−2
30	10	−4	−6	−5	−5	−6	−5	−4	−3	−3
	25	−1	1	0	0	−1	−3	−1	−1	−1
	50	−2	−2	0	−1	0	0	0	−3	1
		Calculation of Power Using Harmonic Mean								
5	10	3	3	0	3	1	1	1	0	1
	25	15	14	14	6	5	6	4	4	2
	50	24	24	22	11	9	8	3	1	2
10	10	1	−1	1	−1	−1	0	−1	0	−2
	25	4	2	2	2	1	1	0	−1	0
	50	7	8	6	1	2	1	1	0	2
15	10	−1	0	1	1	−1	−1	2	−1	−1
	25	−1	−1	4	−1	−1	−1	1	−2	0
	50	4	3	14	1	0	3	0	−1	0
20	10	1	1	2	3	1	1	0	−1	0
	25	−1	0	5	0	−1	−1	1	−1	−2
	50	0	0	4	1	−2	2	−1	−1	0
30	10	5	3	3	1	0	1	−1	0	0
	25	−1	1	0	0	−1	2	−1	0	−1
	50	−1	0	0	0	0	0	0	−2	1

Note. The maximum number of level-1 units per cluster has been restricted to 10, 25, and 50 to condense output. The full table of results is available by request from the first author.