Title:
Design Considerations and Challenges in Planning Complicated Multilevel Experiments

Session Chair:

Nianbo Dong  
University of North Carolina at Chapel Hill  
dong.nianbo@gmail.com

Paper 1:  
Power Analysis of Two- and Three-Level Multisite Moderation Studies

Authors and Affiliations:

Nianbo Dong*  
University of North Carolina at Chapel Hill  
dong.nianbo@gmail.com

Ben Kelcey  
University of Cincinnati  
ben.kelcey@gmail.com

Jessaca Spybrook  
Western Michigan University  
jessaca.spybrook@wmich.edu

Paper 2:  
The Influence of Imbalanced Subgroup Units in Statistical Power for Moderator Effects in Cluster Randomized Trials: Empirical Evidence from IES-Funded Studies

Authors and Affiliations:

Qi Zhang*  
Western Michigan University  
qi.zhang@wmich.edu

Jessaca Spybrook  
Western Michigan University  
jessaca.spybrook@wmich.edu
Paper 3:  
Statistical Power for Mediation Effects in Three-Level Group Randomized Trials

Authors and Affiliations:

Ben Kelcey*  
University of Cincinnati  
ben.kelcey@gmail.com

Kyle Cox  
University of Cincinnati  
coxk5@uc.edu

Paper 4:  
Statistical Power and Optimal Design for Multisite and Cluster-Randomized Studies When the Outcome is not Fully Reliable

Authors and Affiliations:

Kyle Cox*  
University of Cincinnati  
coxk5@uc.edu

Ben Kelcey  
University of Cincinnati  
ben.kelcey@gmail.com

Discussant:

Spyros Konstantopoulos  
Michigan State University  
spyros@msu.edu

*Presenting Author
Symposium Justification
Design Considerations and Challenges in Planning Complicated Multilevel Experiments

Cluster randomized trials (CRTs) and multisite randomized trials (MRTs) have been widely applied to generate rigorous evidence for the main effect, moderation effect, and mediation effects in policy and program evaluation. Statistical power analysis is a critical step in designing CRTs and MRTs to detect educationally meaningful effect sizes or effect size differences. Although extensive research has been done, the statistical tools for power analysis of some complicated CRTs and MRTs are not available. This symposium will introduce recent advances in power analysis of main effects, moderation, and mediation in CRTs and MRTs. Dr. Spyros Konstantopoulos has agreed to serve the discussant.

The first paper presents “Power Analysis of Two- and Three-Level Multisite Moderation Studies”. This paper presents formulas for the statistical power and the minimum detectable effect size difference (MDESD) with confidence intervals for moderator effects in two- and three-level MRTs. This presentation will also demonstrate how to conduct power analysis for multisite moderation studies using the existing free-available software PowerUp!-Moderator.

The second paper presents “The Influence of Imbalanced Subgroup Units in Statistical Power for Moderator Effects in Cluster Randomized Trials: Empirical Evidence from IES-Funded Studies”. This paper will examine the influence of imbalanced subgroup units on the power to detect moderator effects. Specifically, this presentation will report the results of the distribution of cluster-level and individual-level moderator units in CRTs funded by IES and the capacity of these studies to detect moderator effects at both levels.

The third paper presents “Statistical Power for Mediation Effects in Three-Level Group Randomized Trials”. This paper develops closed-form expressions to estimate the statistical power to detect mediation effects in three-level CRTs. This presentation will report the results of a set of formulas and software tools that guide researchers in the planning multilevel studies incorporating mediation. The tests of mediation that can be used both in the planning and analysis phases include the asymptotic Sobel test, component-wise joint test, the resampling-based Monte Carlo interval test, and the partial posterior predictive distribution test.

The fourth paper presents “Statistical Power and Optimal Design for Multisite and Cluster-Randomized Studies When the Outcome is Not Fully Reliable”. This paper derives simple closed-form expressions that detail power and optimal sample allocation for two-level CRTs and MRTs when the outcome is subject to measurement error. This presentation will outline the detrimental effects of outcome measurement error in terms of power and optimal sample allocation and demonstrate power and optimal sampling calculations that account for this error in order to improve study designs.

Collectively, these four papers provide a resource for researchers seeking to design CRTs and MRTs with adequate power to detect the main, moderation, and/or mediation effects of programs and policies to answer questions such as “what works”, “for whom it works and under what conditions it works”, and “why does it work?”.
Paper 1

Power Analysis of Two- and Three-Level Multisite Moderation Studies

Authors and Affiliations:

Nianbo Dong
University of North Carolina at Chapel Hill
116 Peabody Hall, CB 3500
Chapel Hill, NC 27599
Phone: (919)843-9553
dong.nianbo@gmail.com

Ben Kelcey
University of Cincinnati
3311B RECCENTER
Cincinnati, Ohio 45221
Tel: 513-556-3608
ben.kelcey@gmail.com

Jessaca Spybrook
Western Michigan University
3571 Sangren Hall
Kalamazoo, Michigan 49008
Phone: (269) 387-3889
jessaca.spybrook@wmich.edu
Power Analysis of Two- and Three-Level Multisite Moderation Studies

Background and Purposes

Researchers and policy makers are not only interested in the main/average treatment effects but also treatment effect variation such as moderator effects (Weiss, Bloom, & Brock, 2014). For example, it may be that an intervention is more effective in small schools or for low performing students, and that such school or individual characteristics moderate the treatment effect. In planning such a study, a key design consideration is the sample size necessary to achieve adequate statistical power. There exist the tools for power analyses of the main effects of Multisite randomized trials (MRTs) (e.g., Borenstein & Hedges, 2012; Dong & Maynard, 2013; Konstantopoulos, 2008; Raudenbush, Spybrook, Congdon, Liu, & Martinez, 2011) and for power analyses of moderator effects in two- and three-level cluster randomized trails (CRTs) (e.g., Dong, Kelcey, & Spybrook, 2018; Spybrook, Kelcey, & Dong, 2016). Although Bloom and Spybrook (2017) and Raudenbush and Liu (2000) provided a simple framework for power analysis of a binary site-level moderator effect in two-level MRTs, there is no comprehensive statistical tool for power analyses of moderator effects for different types of moderators (e.g., a continuous site-level moderator, a binary or continuous individual level moderator) and different types of models (e.g., random or non-randomly varying slope models). It is still not clear how the intraclass correlations, the covariates, the sample size allocations and the moderators at different levels, and the treatment effect variation/heterogeneity coefficients affect the statistical power of the moderator effects. Given the increasing uses of multisite studies in educational research, the statistical tools and software for power analyses of the effects of moderators at different levels would enhance the capacity of researchers for designing rigorous studies to answer research questions related to the treatment effect variation.

To addresses this gap, in this paper we derive formulas for the statistical power and the minimum detectable effect size difference (MDESD) with confidence intervals for moderator effects in two- and three-level multisite randomized studies. The second goal executes the formulas in the existing free-available software (PowerUp!, Dong & Maynard, 2013).

Theoretical Framework, Methods, and Results

Table 1 below presents the list of two- and three-level multisite moderation designs that are covered in this paper. In designing multisite studies, there are multiple options concerning the levels of the treatment and moderator variables, the random or nonrandomly varying slopes (coefficients) of the treatment variable and the interaction between the treatment and moderator variables, and the distributions of the moderators. The formulas of the power and the minimum detectable effect size difference with confidence intervals for both binary and continuous moderators in Columns 5-8 in Table 1 are to be derived.

[Table 1 about here]

To test for the same and cross-level moderation for Models MRT2-1R-1 and MRT2-1R-2 in Table 1, we use two-level random slope hierarchical linear modeling (HLM) (Raudenbush & Bryk, 2002):

Level 1: \( Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}M_{ij}^{(1)} + \beta_{3j}T_{ij}M_{ij}^{(1)} + \beta_{4j}X_{ij} + r_{ij} \), \( r_{ij} \sim N(0, \sigma_{r}^{2}) \). (1)

Level 2: \( \beta_{0j} = \gamma_{00} + u_{0j} \)
\[ \beta_{1j} = \gamma_{10} + \gamma_{11} M_{ij}^{(2)} + u_{1j} \]
\[ \beta_{2j} = \gamma_{20} \]
\[ \beta_{3j} = \gamma_{30} + u_{3j} \]
\[ \beta_{4j} = \gamma_{40} \]
\[ \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{3j} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01|M} & \tau_{03} \\ \tau_{10|M} & \tau_{11|M} & \tau_{13|M} \\ \tau_{30} & \tau_{31|M} & \tau_{33} \end{pmatrix} \right). \] (2)

\[ Y_{ij} \] is the achievement for student \( i \) in school \( j \). The treatment variable, \( T_{ij} \), is a binary variable indicating whether the student receives the tutoring intervention. \( X_{ij} \) is a level-1 covariate. \( M_{ij}^{(1)} \) is a continuous level-1 moderator, and \( M_{ij}^{(1)} \sim N(0, S_{M}^{(1)}) \) and \( M_{ij}^{(2)} \) is a continuous level-2 moderator, and \( M_{ij}^{(2)} \sim N(0, S_{M}^{(2)}) \). The parameter, \( \gamma_{10} \), estimates the average treatment effect. Of interest for the moderator analysis are the parameters \( \gamma_{30} \) and \( \gamma_{11} \), which represent the treatment effect depending on the moderators \( M_{ij}^{(1)} \) and \( M_{ij}^{(2)} \).

By extending Snijders’ (2001, 2005) work on the variance of the estimated regression coefficients of a level-1 variable with random slope and Raudenbush and Liu’s (2000) work on the power of a binary level-2 moderator effect, we can derive the standardized noncentrality parameters, power and MDESD formulas for level-1 and level-2 moderator effect.

For example, the MDESD regarding Cohen’s \( d \) for a binary level-1 moderator (\( S_{M}^{(1)} = Q_{1}(1 - Q_{1}) \)) is:

\[ MDESD(\hat{\delta}_{1b}) = M_{v} \sqrt{ \frac{(1-R_{2}^{2})(1-\rho)}{jn(1-P)Q_{1}(1-Q_{1})} } \] (3)

where the degrees of freedom is \( J - 1 \).

The 100*(1-\( \alpha \))% confidence interval for \( MDESD(\hat{\delta}_{1b}) \) is given by:

\[ (M_{v} \pm t_{\alpha/2}) \sqrt{ \frac{(1-R_{2}^{2})(1-\rho)}{jn(1-P)Q_{1}(1-Q_{1})} } \] (4)

The MDESD regarding Cohen’s \( d \) for a binary level-2 moderator (\( S_{M}^{(2)} = Q_{2}(1 - Q_{2}) \)) is:

\[ MDESD(\hat{\delta}_{2b}) = M_{v} \sqrt{ \frac{(1-R_{2}^{2})\omega_{1}^{2} \rho}{jQ_{2}(1-Q_{2})} + \frac{(1-R_{1}^{2})(1-\rho)}{jn(1-P)Q_{2}(1-Q_{2})} } \] (5)

where the degrees of freedom is \( J - 2 \).

The 100*(1-\( \alpha \))% confidence interval for \( MDESD(\hat{\delta}_{2b}) \) is given by:

\[ (M_{v} \pm t_{\alpha/2}) \sqrt{ \frac{(1-R_{2}^{2})\omega_{1}^{2} \rho}{jQ_{2}(1-Q_{2})} + \frac{(1-R_{1}^{2})(1-\rho)}{jn(1-P)Q_{2}(1-Q_{2})} } \] (6)

**Illustration**

We implement the formulas in the existing free-available software PowerUp!-Moderator (Dong, Kelcey, Spybrook, & Maynard, 2018). Table 2 illustrate an example of calculating MDESD for binary level-1 and level-2 moderators with random slopes in a two-level multisite randomized trial where students are randomly assigned to receive intervention with each school (site). Suppose that the unconditional intraclass correlation is 0.25, the standardized effect variability of the level-1 moderator across sites is 0.10, the standardized effect variability of the
treatment effect across sites is 0.30, 40% of students are randomly assigned to the treatment group, 50% of students are female, 60% of schools are in urban, the proportion of variance in Level 1 outcome explained by Level 1 covariates, moderator, treatment variable, and interaction is 50%, the proportion of variance for the random treatment slope explained by Level 2 moderator is 10%, and the desired statistical power is 0.80 for a two-tailed test with a type-I error of 0.05, a sample of 40 schools with 20 students in each school will produce the minimum detectable effect size difference (MDES) between girls and boys of 0.264 with the 95% confidence interval of (0.078, 0.450), and the MDES between urban and rural schools of 0.354 with the 95% confidence interval of (0.105, 0.603).

[Table 2 about here]

**Conclusion and Significance**

Although the statistical power and MDES are affected by the collective effects of those design parameters, some basic conclusions can be drawn: the power will increase (MDES will decrease) when the sample sizes (J and n) increase, the proportions of variance explained at levels 1 and 2 (\( R_1^2 \) and \( R_2^2 \)) increase, the effect variability of level-1 moderation decreases or the treatment effect variability decreases. In addition, balanced signs (e.g., \( P = 0.5, Q_1 = Q_2 = 0.5 \)) have bigger power than unbalanced designs, and level-1 moderation has bigger power than level-2 moderation. In summary, the results of this paper will expand the scope and enhance the quality of evidence generated through multisite moderation studies in program evaluation.
References:


Table 1: *PowerUp!-Moderator* to Detect Moderator Effects in Two and Three-Level Multisite Studies: Models and Corresponding Worksheets

<table>
<thead>
<tr>
<th>Number of Total Levels of Clustering</th>
<th>Model Number</th>
<th>Level of Treatment</th>
<th>Level of Moderator</th>
<th>Slope of Treatment</th>
<th>Binary Moderator</th>
<th>Continuous Moderator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MDES Calculation</td>
<td>MDES Calculation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Power Calculation</td>
<td>Power Calculation</td>
</tr>
<tr>
<td>2</td>
<td>MRT2-1N-1</td>
<td>1</td>
<td>1</td>
<td>Nonrandomly Varying</td>
<td>MRT21N_MDES</td>
<td>MRT21Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT21Nc_Power</td>
<td>MRT21Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT2-1N-2</td>
<td>1</td>
<td>2</td>
<td>Nonrandomly Varying</td>
<td>MRT21N_MDES</td>
<td>MRT21Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT21Nc_Power</td>
<td>MRT21Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT2-1R-1</td>
<td>1</td>
<td>1</td>
<td>Random</td>
<td>MRT21R_MDES</td>
<td>MRT21Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT21Rc_Power</td>
<td>MRT21Rc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT2-1R-2</td>
<td>1</td>
<td>2</td>
<td>Random</td>
<td>MRT21R_MDES</td>
<td>MRT21Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT21Rc_Power</td>
<td>MRT21Rc_Power</td>
</tr>
<tr>
<td>3</td>
<td>MRT3-1N-1</td>
<td>1</td>
<td>1</td>
<td>Nonrandomly Varying</td>
<td>MRT31N_MDES</td>
<td>MRT31Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT31Nc_Power</td>
<td>MRT31Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-1N-2</td>
<td>1</td>
<td>2</td>
<td>Nonrandomly Varying</td>
<td>MRT31N_MDES</td>
<td>MRT31Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT31Nc_Power</td>
<td>MRT31Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-1N-3</td>
<td>1</td>
<td>3</td>
<td>Nonrandomly Varying</td>
<td>MRT31N_MDES</td>
<td>MRT31Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT31Nc_Power</td>
<td>MRT31Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-1R-1</td>
<td>1</td>
<td>1</td>
<td>Random</td>
<td>MRT31R_MDES</td>
<td>MRT31Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT31Rc_Power</td>
<td>MRT31Rc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-1R-2</td>
<td>1</td>
<td>2</td>
<td>Random</td>
<td>MRT31R_MDES</td>
<td>MRT31Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT31Rc_Power</td>
<td>MRT31Rc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-1N-3</td>
<td>1</td>
<td>3</td>
<td>Random</td>
<td>MRT31R_MDES</td>
<td>MRT31Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT31Rc_Power</td>
<td>MRT31Rc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-2N-1</td>
<td>2</td>
<td>1</td>
<td>Nonrandomly Varying</td>
<td>MRT32N_MDES</td>
<td>MRT32Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT32Nc_MDES</td>
<td>MRT32Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-2N-2</td>
<td>2</td>
<td>2</td>
<td>Nonrandomly Varying</td>
<td>MRT32N_MDES</td>
<td>MRT32Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT32Nc_MDES</td>
<td>MRT32Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-2N-3</td>
<td>2</td>
<td>3</td>
<td>Nonrandomly Varying</td>
<td>MRT32N_MDES</td>
<td>MRT32Nc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT32Nc_MDES</td>
<td>MRT32Nc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-2R-1</td>
<td>2</td>
<td>1</td>
<td>Random</td>
<td>MRT32R_MDES</td>
<td>MRT32Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT32Rc_MDES</td>
<td>MRT32Rc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-2R-2</td>
<td>2</td>
<td>2</td>
<td>Random</td>
<td>MRT32R_MDES</td>
<td>MRT32Rc_MDES</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRT32Rc_MDES</td>
<td>MRT32Rc_Power</td>
</tr>
<tr>
<td></td>
<td>MRT3-2R-3</td>
<td>2</td>
<td>3</td>
<td>Random</td>
<td>MRT32R_MDES</td>
<td>MRT32Rc_MDES</td>
</tr>
</tbody>
</table>
Table 2: *PowerUp!-Moderator* Illustration: MDESD Calculator for Two-Level Multisite Randomized Trials — Treatment at Level 1 and Binary Moderators at Level-1 and 2 (Random slope model)

<table>
<thead>
<tr>
<th>Model MRT21R: MDESD Calculator for Two-Level Multisite Randomized Trials — Treatment at Level 1 and Binary Moderators at Level-1 and 2 (Random slope model)</th>
<th>Assumptions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha Level (α)</td>
<td>0.05</td>
<td>Probability of a Type I error</td>
</tr>
<tr>
<td>Two-tailed or One-tailed Test?</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Power (1-β)</td>
<td>0.80</td>
<td>Statistical power (1-probability of a Type II error)</td>
</tr>
<tr>
<td>Rho (ICC)</td>
<td>0.25</td>
<td>Proportion of variance among sites: $\rho = \tau_{00}^2/(\tau_{00}^2 + \sigma^2)$</td>
</tr>
<tr>
<td>$\omega_3$</td>
<td>0.10</td>
<td>The standardized effect variability of the level-1 moderation across sites: $\omega_3 = \tau_{33}^2/\tau_{40}^2$</td>
</tr>
<tr>
<td>$\omega_4$</td>
<td>0.30</td>
<td>The standardized effect variability of the treatment effect across sites: $\omega_4 = \tau_{11}^2/\tau_{40}^2$</td>
</tr>
<tr>
<td>$P$</td>
<td>0.40</td>
<td>Proportion of Level 1 units randomized to treatment: $n_T/(n_T + n_C)$</td>
</tr>
<tr>
<td>$Q_1$</td>
<td>0.50</td>
<td>Proportion of Level 1 units in Moderator subgroup: $n_1/(n_1 + n_0)$</td>
</tr>
<tr>
<td>$Q_2$</td>
<td>0.60</td>
<td>Proportion of Level 2 units in Moderator subgroup: $J_1/(J_1 + J_0)$</td>
</tr>
<tr>
<td>$R_1^2$</td>
<td>0.50</td>
<td>Proportion of variance in Level 1 outcome explained by Level 1 covariates, moderator, treatment variable, and interaction.</td>
</tr>
<tr>
<td>$R_2^2$</td>
<td>0.10</td>
<td>Proportion of variance for the random treatment slope explained by Level 2 moderator.</td>
</tr>
<tr>
<td>$\bar{n}$ (Average Sample Size for Level 1)</td>
<td>20</td>
<td>Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)</td>
</tr>
<tr>
<td>$\bar{J}$ (Average Sample Size for Level 2)</td>
<td>40</td>
<td>Total number of Level 2 units (sites)</td>
</tr>
<tr>
<td>MDESD($\hat{\delta}_{1b}$)</td>
<td>0.264</td>
<td>Minimum Detectable Effect Size Difference regarding Cohen's $d$ for Level-1 Moderator</td>
</tr>
<tr>
<td>95% Confidence Interval</td>
<td>(0.078, 0.45)</td>
<td>95% Confidence Interval of MDESD($\hat{\delta}_{1b}$)</td>
</tr>
<tr>
<td>MDESD($\hat{\delta}_{1a}$)</td>
<td>0.354</td>
<td>Minimum Detectable Effect Size Difference regarding Cohen's $d$ for Level-2 Moderator</td>
</tr>
<tr>
<td>90% Confidence Interval</td>
<td>(0.105, 0.603)</td>
<td>95% Confidence Interval of MDESD($\hat{\delta}_{2a}$)</td>
</tr>
</tbody>
</table>

Note: The parameters in yellow cells need to be specified. The MDESD will be calculated automatically.
Paper 2

Title: The Influence of Imbalanced Subgroup Units in Statistical Power for Moderator Effects in Cluster Randomized Trials: Empirical Evidence from IES-Funded Studies

Authors and Affiliations:

Qi Zhang*
Western Michigan University
qi.zhang@wmich.edu

Jessaca Spybrook
Western Michigan University
ejessaca.spybrook@wmich.edu
The Influence of Imbalanced Subgroup Units in Statistical Power for Moderator Effects in Cluster Randomized Trials: Empirical Evidence from IES-Funded Studies

**Background:** Cluster Randomized Trials (CRTs) with schools as the unit of random assignment are often implemented to assess the efficacy of educational interventions in K-12 settings. An a-priori power analysis is conducted in the planning stage of a CRT to determine the capacity of the study design to sufficiently detect a meaningful treatment effect. With the advancement of methods and tools for calculating statistical power, partially driven by the demand from the Institute of Educational Sciences (IES) for rigorous designs, we have seen an increase in the capacity of IES-funded impact studies to detect meaningful main treatment effects (Spybrook, Shi, & Kelcey, 2016). As more studies are becoming equipped to answer the “does the program work” question, there is also an interest in the capacity of studies to answer questions regarding treatment effect heterogeneity, such as “for whom” and “under what conditions” programs work. Moderator analyses are often conducted to answer questions regarding treatment effect heterogeneity. Power calculations for moderator effect are often done with the assumption that the proportion of cluster- or individual-level units are equal in moderator subgroups, that is the number units within moderator subgroups are the same in treatment and control conditions. However, when there is an imbalance in the subgroup units, the effective sample size to detect the moderator effect becomes smaller and the power associated with detecting these effects decreases (Spybrook et al., 2016).

**Purpose:** The purpose of this study is to examine the influence of imbalanced subgroup units on the power to detect moderator effects. Specifically, we are interested in the distribution of cluster-level and individual-level moderator units in CRTs funded by IES and the capacity of these studies to detect moderator effects at both levels. The goal is to derive an understanding of the state of the field of education research regarding powering studies to detect moderator effects. With this knowledge, we hope to suggest design elements that can help studies to expand their capacity to answer more questions regarding the heterogeneity of treatment effect.

**Method:** We collected information regarding school characteristics for 22 Goal-3 and Goal-4 CRTs funded by IES between 2005 and 2016. The studies in our sample included CRTs focused on Grade K-12 that randomized schools or teachers. We interviewed PIs to learn more about their recruiting practice and collected information about the schools in their studies. Then from the Common Core of Data, we gathered information on potential moderators, such as school’s status in implementing Title I, school location, proportion of students in the free/reduced-lunch (FRL) program, and proportion of minority students. Currently, we have collected information for approximately 1,000 schools.

The moderator variables were dichotomous or categorical. For categorical variables, we dichotomized them based their mean values in the population. For instance, an average of 64% of students within the schools in our population dataset qualified for FRL. Therefore, schools with more than 64% of students in FRL programs were considered as “low socioeconomic status (SES)” and vice versa. The proportion of schools with low SES were subsequently calculated for each study. Table 1 shows the moderator distribution for a sample of 7 studies and 4 selected school-level moderators. For the presentation, we will expand to all 22 studies and include additional moderator variables at both the school and student-level.
Table 1. School-level moderator distribution for a sample of 7 studies.

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Study 1 (J = 30)</th>
<th>Study 2 (J = 36)</th>
<th>Study 3 (J = 38)</th>
<th>Study 4 (J = 40)</th>
<th>Study 5 (J = 46)</th>
<th>Study 6 (J = 49)</th>
<th>Study 7 (J = 70)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>0.53</td>
<td>0.09</td>
<td>0.30</td>
<td>0.23</td>
<td>0.70</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>High Minority*</td>
<td>0.35</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.07</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>Low SES**</td>
<td>0.39</td>
<td>0.66</td>
<td>0.47</td>
<td>0.48</td>
<td>1.00</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>Schoolwide Title I</td>
<td>0.69</td>
<td>0.72</td>
<td>0.47</td>
<td>0.48</td>
<td>0.67</td>
<td>0.97</td>
<td>0.79</td>
</tr>
</tbody>
</table>

* High Minority: > 67% non-White students.
**Low SES: > 64% student in FRL programs.

For the purposes of this proposal, we calculated statistical power to detect a cluster level moderator effect, represented by the minimum detectable effect size difference (MDESD) for the 7 studies of interventions aimed to improve student achievement. In the full paper we will also include calculations for individual-level moderators. The power calculations were based on the 2-level Hierarchical Linear Model (HLM) with students nested within schools, and treatment and moderator at the school-level. For studies with a 3-level CRT design, we ignored the teacher/classroom-level for this calculation, which is not expected to influence the power calculations (Zu et al., 2012). Table 2 below shows the design parameters used for this analysis, which were based on those reported by Hedges and Hedberg (2007) for student achievement outcomes. We used an upper and lower bound for the ICC and school-level $R^2$, and a single value for the individual-level $R^2$. Different combinations of these parameters led to a range of MDESDs. Power calculations were operationalized in the program PowerUp!-Moderator (Dong et al., 2016).

Table 2. Empirical estimates used for MDESD calculations*.

<table>
<thead>
<tr>
<th>Student Achievement Outcome</th>
<th>School-level ICC</th>
<th>School-level $R^2$</th>
<th>Individual-level $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.15, 0.25</td>
<td>0.50, 0.80</td>
<td>0.40</td>
</tr>
</tbody>
</table>

*Calculations were based on these additional assumptions: two-tailed test, alpha = 0.05, equal allocation at all levels.

**Result:** Figure 1 shows the MDESD for the sample of 7 studies and four moderators. In each graph, studies were arranged in the order from the smallest to largest number of total schools (J) in the study. Within each graph, the black lines represented MDESD calculated based on population proportions. For comparison purpose, we calculated MDESD based on equal distribution of moderator units for each study, which was represented in red lines.
Figure 1. Calculated MDESD based on population proportion (black line) and balanced proportions (red line) for the following moderators: a) Urbanicity, b) High Minority, c) Low SES, d) Schoolwide Title I. Note that MDESD could not be calculated for studies with moderator distribution of either 0 or 1, as they indicated that there were no variations in the moderator. Based on this, 5 out of the 7 studies were not designed to examine the treatment effect variation across minority status. Similarly, 2 studies and 1 study could not detect moderation effect based on SES and urbanicity, respectively.

The figure shows how the distributions of moderator units impacts power. For example, in Figure 1a the imbalance in the distribution of urbanicity makes the MDESD much larger than in the balanced case. Further, we can see the influence of the distribution of the moderator as it relates to the sample sizes. That is, in Figure 1a study 1 was designed to detect a smaller MDESD than study 3, even though study 1 had less number of schools.

Conclusion: In general, powering studies to detect meaningful moderator effects is challenging. The distribution of moderators for many studies creates further complications as those with proportions equal or close to 0 or 1 for a particular moderator can greatly increase the MDESD. Hence, if moderator effects are a primary component of the study, the distribution of the moderators should be considered in the planning phase. The full paper will also include results for individual moderator effects.
References


Statistical Power for Mediation Effects in Three-Level Group Randomized Trials

Authors and Affiliations:
Ben Kelcey
University of Cincinnati
ben.kelcey@gmail.com

Kyle Cox
University of Cincinnati
coxk5@uc.edu
Statistical Power for Mediation Effects in Three-Level Group Randomized Trials

Purpose

We develop closed-form expressions to estimate the statistical power to detect mediation effects in three-level group-randomized trials. Mapping the sensitivity of three-level designs is of critical importance in education because such design sensitivity directly governs the types of evidence researchers can bring to bear on theories of action under common sample sizes. The results provide a set of formulas and software tools that guide researchers in the planning multilevel studies incorporating mediation. We extend tests of mediation that can be used both in the planning and analysis phases including the asymptotic Sobel test, component-wise joint test, the resampling-based Monte Carlo interval test, and the partial posterior predictive distribution test. These tests represent classical and modern approaches, capture important differences among tests in terms type one error rates and power levels, can each be implemented in the design phase, and are collectively representative of the range of tests most commonly used in the literature.

Background

Three-level designs are quite common in education research because they align well with the core organizational structure of schooling—students nested within classrooms nested within schools (e.g., Spybrook, Shi, & Kelcey, 2016). An important design consideration in three-level designs is the power with which we can detect effects if they exist. Historically, studies have been exclusively designed with a focus on detecting main effects. Recent literature has, however, expanded that consideration to include mediation effects because the comprehensive study of the components of a theory of action is critical to advancing scientific theories (e.g., Desimone, 2009; US DoE & NSF, 2013). Although prior literature has detailed the power to detect main and moderation effects in three-level designs, literature regarding the calculation of power for mediation in three-level designs has been largely absent (Raudenbush, 1997; Dong, Kelcey, & Spybrook, 2018).

Model

For brevity, we outline the development and application of power analyses using one simple (albeit sub-optimal) example—the Sobel test with students nested within classrooms nested within schools. The Sobel test has received due criticism for shortcomings; however, for the purposes of an abstract, the simplicity of its derivation provides a good conceptual description of the more general analyses and results. We again note that although we outline the Sobel test under a 3-2-1 mediation example, our full study derives power formulas for the aforementioned tests and for a broader range multilevel mediation effects as detailed through the potential outcomes framework (e.g., Kelcey, Dong, Spybrook, & Cox, 2017).

To conceptually outline our analysis, we consider one example estimand—the mediation effect that describes the extent to which a school-assigned treatment ($T$) impacts an individual-level outcome ($Y$) through a classroom-level mediator ($M$) conditional upon covariates. The corresponding path model is
Mediator model
\[
M_{jk} = \pi_{0k} + \pi_1 (\bar{X}_{jk} - \bar{X}_k) + \pi_2 (W_{jk} - \bar{W}_k) + \pi_3 U_{jk} + \varepsilon_{jk}^M
\]
\[
\varepsilon_{jk}^M \sim N(0, \sigma_{M_{jk}}^2)
\]
\[
\pi_{0k} = \zeta_{00} + aT_k + \zeta_{01} \bar{X}_k + \zeta_{02} \bar{W}_k + \zeta_{03} Z_k + u_{0k}^M
\]
\[
u_{0k}^M \sim N(0, \tau_{M_{jk}}^2)
\]

Outcome model
\[
Y_{ijk} = \beta_0ijk + \beta_1 (X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon_{ijk}^Y
\]
\[
\varepsilon_{ijk}^Y \sim N(0, \sigma_{Y_{ijk}}^2)
\]
\[
\beta_0ijk = \gamma_{00} + \gamma_{01} (M_{jk} - \bar{M}_k) + \gamma_{02} (X_{ijk} - \bar{X}_k) + \gamma_{03} (W_{ijk} - \bar{W}_k) + \gamma_{04} U_{ijk} + \nu_{ijk}^Y
\]
\[
\nu_{ijk}^Y \sim N(0, \tau_{Y_{ijk}}^2)
\]

Here we use \( Y_{ijk} \) as the outcome for student \( i \) in class \( j \) in school \( k \), and add \( V_{ijk} \) as a student-level covariate that varies only among students, \( M_{jk} - \bar{M}_k \) as the group-centered class-level mediator with coefficient \( \beta_2 \), \( \bar{M}_k \) as the mean of the mediator in school \( k \) with path coefficient \( B \), \( c' \) as the treatment-outcome conditional path coefficient, and \( \nu_{0ijk}, \nu_{ijk}^Y, \nu_{ijk}^Y \) as the school, class, and student error terms. Given this formulation, our example analysis investigates the power to detect the mediation effect defined by \( aB \) (we assume no interactions for ease of presentation).

One of the most common tests of statistical significance for indirect effects is the Sobel test that compares the ratio of the product of the \( a \) and \( B \) coefficients to the standard error \( (\sigma_{ab}) \) of this product
\[
z_{Sobel} = \frac{aB}{\sqrt{\sigma_{ab}^2}}
\]

Our omitted derivations extend this test to three-level settings such that the resulting Sobel test for a mediation effect is
\[
z_{Sobel}^{\text{SL}} = \frac{aB}{\sqrt{\sigma_{ab}^2}} = \frac{aB}{\sqrt{\frac{\tau_{M_{jk}}^2 (1 - R_{M_{jk}}^2) + (1 - R_{MT_{jk}}^2) \sigma_{M_{jk}}^2 / n_z}{n_z \sigma_{M_{jk}}^2 (1 - R_{M_{jk}}^2)}}}
\]

We use \( \tau_{M_{jk}}^2 \) and \( \sigma_{M_{jk}}^2 \) as the unconditional school- and class-level variances of the mediator, \( \sigma_Y^2 \) as variance of the treatment, \( R_{M_{jk}}^2 \) and \( R_{MT_{jk}}^2 \) as the school- and class-level mediator variance explained by predictors in the mediation model, \( R_{T_{jk}}^2 \) as treatment variance explained by predictors, and \( n_z \) and \( n_2 \) as the school- and class-level sample sizes. Similarly we introduce \( \nu_{Y_{ijk}}, \tau_{Y_{ijk}}, \sigma_{Y_{ijk}}^2, \tau_{MT_{jk}}^2, \sigma_{MT_{jk}}^2 \) as the unconditional school- and class-level treatment-by-mediator variance and \( R_{MT_{jk}}^2 \), \( R_{MT_{jk}}^2 \), and \( R_{MT_{jk}}^2 \) as the school- and class-level interaction and mediator variances explained by other predictors in the outcome model.

Assuming the alternative hypothesis is true, the power of a two-sided test to detect the multilevel mediation effect can be approximated with
\[
P(z_{Sobel} > z_{\text{critical}}) = 1 - \Phi(z_{\text{critical}} - z_{Sobel}^{\text{SL}}) + \Phi(-z_{\text{critical}} - z_{Sobel}^{\text{SL}})
\]
where $\Phi$ is the normal distribution with $z_{\text{critical}}$ as the chosen critical value (e.g., 1.96) corresponding to a nominal type one error rate.

**Findings**

To illustrate the utility of the results, consider an example in which schools are randomly assigned to treatment conditions (e.g., professional development), we observe a classroom-level mediator (e.g., teacher instructional quality), and record a student-level outcome (achievement). Assume that the theory of action suggests that exposure to professional development (treatment) improves student achievement (outcome) by improving on teacher instruction (mediator). In planning a study, we might inquire as to approximately how many schools we need to have a reasonably high chance of detect the mediation effect if it exists. Previously, formulas were not available to determine such sample sizes and estimate power. Using the resulting formulas, let us assume we anticipate the following parameter values for our study:

- $\alpha = 0.50$ (treatment-mediator relationship [Cohen’s $d$ scale])
- $B = 0.30$ (mediator-outcome relationship [Standardized regression scale])
- $\nu_y^2 = 0.10$ (unconditional outcome variance at school-level)
- $\tau_y^2 = 0.10$ (unconditional outcome variance at class-level)
- $\sigma_y^2 = 0.80$ (unconditional outcome variance at individual-level)
- $\tau_M^2 = 0.20$ (unconditional outcome variance at school-level)
- $\sigma_M^2 = 0.80$ (unconditional outcome variance at class-level)
- $R^2_{y_0} = R^2_{y_1} = R^2_{y_2} = 0.50$ (outcome variance explained at each level)
- $R^2_{M_0} = R^2_{M_1} = 0.50$ (mediator variance explained at each level)
- $P = 0.50$ (proportion of schools receiving treatment)
- $n_s = 4$ (classrooms/school)
- $n_t = 20$ (students/classroom)

The resulting power curve as a function of the school sample size is plotted in Figure 1. Based on our derivations, we would expect that sampling about 40 schools would yield about a 0.80 chance of detecting the mediation effect under a three-level group-randomized design.

**Conclusions**

Designing studies with the capacity to test whether or not a program works and to examine the mechanisms underlying the program theory has become a prominent and critical aim of research studies. To date there is little guidance for designing studies to detect three-level mediation effects. Our work bridges that gap by establishing a flexible framework from which to estimate power under many common three-level designs and implements them in easy to use software. With these tools, we hope to streamline multilevel mediation power analyses in ways that help researchers understand how to carefully plan studies.
References


Figure 1
Power to detect mediation effect in a three-level group-randomized design
Statistical Power and Optimal Design for Multisite and Cluster-Randomized Studies When the Outcome is not Fully Reliable

Authors and Affiliations:

Kyle Cox (presenting)
University of Cincinnati
coxk5@uc.edu

Ben Kelcey
University of Cincinnati
ben.kelcey@gmail.com
Statistical Power and Optimal Design for Multisite and Cluster-Randomized Studies When the Outcome is not Fully Reliable

Purpose

The purpose of this study was to derive simple closed-form expressions that detail power and optimal sample allocation for two-level multisite and cluster-randomized studies (MSRTs and CRTs) when the outcome is subject to measurement error. Recent literature has reported a dramatic improvement in the statistical power with which studies in education are conducted (e.g., Spybrook, Shi, & Kelcey, 2016). However, a critical limitation of many power analyses is their disregard of potentially deleterious effects of measurement error (illustrated in Figure 1). More specifically, most studies in education are conducted using latent outcome variables that are subject to measurement error (e.g., mathematics achievement); however, most power analyses are conducted without accounting for the possibility of measurement error. We outline the detrimental effects of outcome measurement error in terms of power and optimal sample allocation and develop power and optimal sampling formulas that account for this error in order to improve study designs. Due to space constraints we outline only CRTs but note that our work also covers multisite designs.

Background

In educational research, it is extremely common to investigate outcomes that are not directly observed and subject to measurement error. Despite the widespread use of latent outcomes, literature has not clearly detailed simple expressions to estimate power and identify optimal sampling allocations in the presence of outcome measurement error.

Method

To frame our study, consider a recent example evaluating the effectiveness of the Teacher-Led Math Inquiry intervention on student mathematics achievement (Hull, et al., 2018). For this two-level CRT, the outcome can be estimated using the following series of hierarchical linear models (Raudenbush & Bryk, 2002).

\[
Y_{ij} = \pi_{0j} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2_Y) \\
\pi_{0j} = \zeta_{00} + \delta T_j + u_{0j} \quad u_{0j} \sim N(0, \tau^2_{\pi})
\]

Here, the outcome (i.e., mathematics achievement) of student \(i\) in school \(j\) is represented by \(Y_{ij}\) which is composed of the mean school outcome \(\pi_{0j}\), and an individual error term \(\varepsilon_{ij} \sim N(0, \sigma^2_Y)\). At the school level, the mean school outcome is decomposed beginning with \(\zeta_{00}\), the overall mean outcome. The relationship between the school-level treatment (i.e., Math Inquiry intervention) and student-level outcome (mathematics achievement) is captured by \(\delta\) as it is associated with the treatment indicator \(T_j\).
Measurement error. We consider the use of a fallible measure of the outcome \( \hat{Y}_j \) rather than the true outcome \( Y_j \). In our example, the mathematics assessment is not a perfectly reliable measure of a student mathematics achievement. We assume that the observed outcome contains non-differential independent measurement errors that have constant variance, a mean of zero and are uncorrelated with other students in the same school or the true values. The observed outcome can be expressed as
\[
\hat{Y}_j = Y_j + e_j, \quad e_j \sim N(0, \sigma_e^2)
\]  
(2)

with \( \hat{Y}_j \) as the observed outcome value for student \( i \) in school \( j \) and \( e_j \) as the independent and normally distributed error with variance \( \sigma_e^2 \). The observed outcome remains normally distributed with individual-level variance \( \sigma_y^2 \) such that \( \sigma_y^2 = \sigma_i^2 + \sigma_e^2 \).

In turn, our analyses probe CRTs with fallibly measured outcomes where the \( \sim \) symbol is used to indicate values based on the observed outcome rather than the true outcome.

Error variance. To develop power and optimal sampling formulas that account for measurement error, we first derive the error variance of the treatment effect when the outcome is subject to measurement error. When the outcome is measured with error, the resulting error variance becomes
\[
\sigma^2_\delta = \frac{\tau^2_Y \lambda^{li}_i + (\sigma_i^2 / \lambda^{li}_i)}{n_1} \frac{1}{n_2 \sigma_e^2}
\]  
(3)

with \( \lambda^{li}_i \) as the individual-level reliability of the outcome (i.e., \( \lambda^{li}_i = \frac{\sigma_y^2}{\sigma_i^2 + \sigma_e^2} \)).

Power. To test inferences regarding the null hypothesis of the treatment effect \( H_0 : \delta = 0 \) we used the \( t \)-distribution with
\[
t_{\delta} = \frac{\bar{\delta}}{\hat{\sigma}_{\delta}} = \frac{\bar{\delta}}{\sqrt{\frac{\tau^2_Y \lambda^{li}_i + (\sigma_i^2 / \lambda^{li}_i)}{n_1} / \sigma_e^2}}
\]  
(4)

Under the alternative hypothesis \( H_1 : \bar{\delta} > 0 \), \( t_{\delta} \) follows a non-central \( t \)-distribution with \( t_{\delta} \) as the non-centrality parameter and \( df = n_2 - 2 \). Power is then determined with
\[
P(|t_{\delta} | > t_{n_2-2,\alpha/2}) = 1 - t(t_{n_2-2,\alpha/2} - t_{\delta}) + t(-t_{n_2-2,\alpha/2} - t_{\delta})
\]  
(5)

where \( t_{n_2-2,\alpha/2} \) is the critical value, the probability of a type-I error is \( \alpha = 0.05 \), and \( t \) represents the cumulative \( t \) density function with appropriate degrees of freedom.

Optimal Sample Allocation. To identify the optimal sampling plan in the presence of measurement error, we begin with the conventional linear cost formulation (Raudenbush, 1997) such that
\[
T = c_x n_2 + c_i n_2 n_1
\]  
(6)
$T$ is the total funds available for a study and $c_1$ and $c_2$ are sampling costs for individuals and groups. Each is typically measured in monetary units.

Incorporating the cost formulation into $\sigma_\delta^2$ and minimizing this expression in terms of $n_1$ produces the optimal sample of individuals per cluster when the outcome is measured with $(\sigma_i^2)$

$$n_i^{opt} = \sqrt{ \frac{c_2}{c_1} \frac{\sigma_i^2 + \sigma_c^2}{r_{yi}}}.$$  

(7)

The resulting equation shows that outcome measurement error will typically increase the optimal number of individuals per group to sample as compared to the assumption that the outcome is fully reliable.

**Results**

We probed the power and $n_i^{opt}$ expressions above to highlight the consequences of ignoring measurement error in CRT planning. Table 1 summarizes study conditions while Tables 2 and 3 present results indicating power rates fall and optimal individual sample size increases as measurement error increases. Differences across conditions are illustrated in Figures 2 and 3.

Our Math Inquiry intervention study provides an illustrative example. Assume we conducted a similar study to Hull et al. (2018) but employed a new observational system to assess student mathematics ability. If we disregarded measurement error in the planning stages of the study, but planned for $n_2 = 280$, $n_1 = 10$, $\sigma_i^2 = .72$, $r_{yi} = .28$, and $\delta = 0.2$, we would achieve a power rate of $\approx .80$.When we include measurement error resulting in $\lambda_{yi}^{L1} \approx 0.56$, our power rate drops to $\approx .74$. We could also plan this study using the optimal sample allocation framework to determine $n_1$ and $n_2$. With $c_2 / c_1 = 100$, $n_i^{opt} \approx 16$ but when we consider outcome unreliability caused by measurement error $n_i^{opt} \approx 21$.

**Significance**

Unobservable latent outcomes that incur some type of measurement error are popular in educational research and all but ensure unreliability in outcomes of multilevel educational experiments. It is crucial for researchers to incorporate this measurement error in the planning stages of these studies. Ignoring measurement error while planning educational experiments can not only lead to inefficiencies but can also result in underpowered studies.
References


TABLE 1.  
*Parameter Values for Investigating the Influence of Measurement Error on Power, Minimum Detectable Effect Size, and Optimal Sample Allocation in Two-level Cluster-randomized Trials*

| $\sigma_e^2$ | $\delta$ | $n_2$ | $n_1$ | $\sigma_y^2$ | $\tau_{y|1}^2$ | $c_2 / c_1$ |
|--------------|----------|-------|-------|--------------|----------------|--------------|
| 0-0.9        | 0.05     | 40    | 10    | 0.9          | 0.1            | 2/1          |
| 0.1          | 80       | 20    |       | 0.7          | 0.3            | 10/1         |
| 0.15         | 40       |       | 0.5   | 0.5          |                | 100/1        |
| 0.25         |          | 40    |       | 0.5          |                | 1000/1       |
| 0.5          |          |       |       |              |                |              |

*Note.* The total budget ($T$) was set as a function of the group to individual cost ratio such that $T = 50(c_2 / c_1)$ and the coefficient representing the treatment effect ($\delta$) was set at 0.5. $\delta$ is not applicable for the MDE investigation, cost ratio ($c_2 / c_1$) is only applicable to the investigation under the optimal sample allocation framework while sample size ($n_1$ and $n_2$) are not applicable for the optimal sample allocation investigation.
TABLE 2.
Power to Detect a Treatment Effect When the Outcome is Measured with Error in a Two-level Cluster-randomized Trial

<table>
<thead>
<tr>
<th>Power</th>
<th>Delta $\delta = 0.25$</th>
<th>Delta $\delta = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_e^2$</td>
<td>0.0</td>
<td>0.25</td>
</tr>
<tr>
<td>Reliability ($\lambda_{11}$)</td>
<td>0.00</td>
<td>0.40</td>
</tr>
<tr>
<td>0.0</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>0.1</td>
<td>0.90</td>
<td>0.50</td>
</tr>
<tr>
<td>0.2</td>
<td>0.82</td>
<td>0.70</td>
</tr>
<tr>
<td>0.3</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td>0.4</td>
<td>0.69</td>
<td>0.60</td>
</tr>
<tr>
<td>0.5</td>
<td>0.64</td>
<td>0.50</td>
</tr>
<tr>
<td>0.6</td>
<td>0.60</td>
<td>0.40</td>
</tr>
<tr>
<td>0.7</td>
<td>0.56</td>
<td>0.30</td>
</tr>
<tr>
<td>0.8</td>
<td>0.53</td>
<td>0.20</td>
</tr>
<tr>
<td>0.9</td>
<td>0.50</td>
<td>0.10</td>
</tr>
<tr>
<td>0.0</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>


Note. In this case $\sigma_y^2 = 0.9$ and $\tau_i^2 = 0.1$ and % Power loss refers to the decrease in power from an outcome with full reliability to 50% outcome reliability.
### Table 4. Influence of Measurement Error in the Outcome Under the Optimal Sample Allocation Framework on Individual Sample Size, Power, and Minimum Detectable Effect Size in Two-level Cluster-randomized Trials

<table>
<thead>
<tr>
<th>Error ($\sigma_e^2$)</th>
<th>Reliability ($\lambda_{\text{LL}}$)</th>
<th>Cost Ratio: 10</th>
<th></th>
<th>Cost Ratio: 100</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$n_i^{\text{opt}}$</td>
<td>Power</td>
<td>MD E</td>
<td>Power</td>
</tr>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>9.487</td>
<td>0.26</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>0.10</td>
<td>0.90</td>
<td>10.00</td>
<td>0.25</td>
<td>0.01</td>
<td>0.52</td>
</tr>
<tr>
<td>0.20</td>
<td>0.82</td>
<td>10.48</td>
<td>0.24</td>
<td>0.02</td>
<td>0.53</td>
</tr>
<tr>
<td>0.30</td>
<td>0.75</td>
<td>10.95</td>
<td>0.23</td>
<td>0.03</td>
<td>0.55</td>
</tr>
<tr>
<td>0.40</td>
<td>0.69</td>
<td>11.40</td>
<td>0.22</td>
<td>0.04</td>
<td>0.56</td>
</tr>
<tr>
<td>0.50</td>
<td>0.64</td>
<td>11.83</td>
<td>0.21</td>
<td>0.05</td>
<td>0.57</td>
</tr>
<tr>
<td>0.60</td>
<td>0.60</td>
<td>12.24</td>
<td>0.21</td>
<td>0.05</td>
<td>0.58</td>
</tr>
<tr>
<td>0.70</td>
<td>0.56</td>
<td>12.64</td>
<td>0.20</td>
<td>0.06</td>
<td>0.59</td>
</tr>
<tr>
<td>0.80</td>
<td>0.53</td>
<td>13.03</td>
<td>0.19</td>
<td>0.07</td>
<td>0.60</td>
</tr>
<tr>
<td>0.90</td>
<td>0.50</td>
<td>13.41</td>
<td>0.19</td>
<td>0.07</td>
<td>0.61</td>
</tr>
</tbody>
</table>

| % Change | 41.4 | -28.2 | 21.3 | 41.4 | -14.0 | 9.9 |

*Note.* In this selected condition $\delta = 0.25$ and $\sigma_i^2 = 0.9 \tau_{ij} = 0.1$. 

29
FIGURE 1. Power to detect the main effect in a two-level cluster randomized study as a function of cluster sample size when individual-to-cluster-variance in the outcome is 0.9:0.1, the magnitude of the treatment effect is 0.25, the individual per cluster sample size is 10, and there is no measurement error in the outcome (solid) and measurement error in the outcome causing outcome reliability to be .5 (dash).
FIGURE 2

FIGURE 1. Power to detect a main effect in a two-level cluster randomized study as a function of measurement error in the level-one outcome when individual to cluster variance is 0.9, 0.1, delta is 0.05, 0.1, 0.15, 0.25, and 0.5 and the individual sample size is 10 (solid), 20 (dashed), and 40 (solid) with group sample size of 40 (thin lines) and 80 (thick lines).
FIGURE 2. Optimal individual sample size for two-level cluster randomized studies as a function of measurement error in the level-one outcome when individual to cluster level variance is 0.9:0.1, 0.7:0.2, and 0.5:0.5 and the group-to-individual cost ratio is 100:1 (dashed), 100:1 (dot), 10:1 (dashdot), and 2:1 (solid).