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: $Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}M_{ij}^{(1)} + \beta_{3j}T_{ij}M_{ij}^{(1)} + \beta_{4j}X_{ij} + r_{ij}, \;\; r_{ij} \sim N(0, \sigma^2_{|T,M,X}).$ (1)

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

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

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 $M_{ij}^{(1)} \sim N(0, S_{M^{(1)}}^2)$ and $M_j^{(2)}$ is a continuous level-2 moderator, and $M_j^{(2)} \sim N(0, S_{M^{(2)}}^2)$. The parameter, γ_{10}, estimates the average treatment effect. Of interest for the moderator analysis are the parameters γ_{30} and γ_{11}, which represent the treatment effect depending on the moderators $M_{ij}^{(1)}$ and $M_j^{(2)}$.

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 derived 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 ($S_{M^{(1)}}^2 = Q_1(1-Q_1)$) is:

$$MDESD\left(\left|\hat{\delta}_{1b}\right|\right) = M_v\sqrt{\frac{\omega_3\rho}{J} + \frac{(1-R_1^2)(1-\rho)}{JnP(1-P)Q_1(1-Q_1)}}, \quad (3)$$

where the degrees of freedom is $J - 1$.

The 100*(1-α)% confidence interval for $MDESD\left(\left|\hat{\delta}_{1b}\right|\right)$ is given by:

$$(M_v \pm t_{\alpha/2})\sqrt{\frac{\omega_3\rho}{J} + \frac{(1-R_1^2)(1-\rho)}{JnP(1-P)Q_1(1-Q_1)}} \ . \quad (4)$$

The MDESD regarding Cohen's *d* for a binary level-2 moderator ($S_{M^{(2)}}^2 = Q_2(1-Q_2)$) is:

$$MDESD\left(\left|\hat{\delta}_{2b}\right|\right) = M_v\sqrt{\frac{(1-R_2^2)\omega_1\rho}{JQ_2(1-Q_2)} + \frac{(1-R_1^2)(1-\rho)}{JnP(1-P)Q_2(1-Q_2)}}, \quad (5)$$

where the degrees of freedom is $J - 2$.

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

$$(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R_2^2)\omega_1\rho}{JQ_2(1-Q_2)} + \frac{(1-R_1^2)(1-\rho)}{JnP(1-P)Q_2(1-Q_2)}} \ . \quad (6)$$

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 moderation 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 (MDESD) between girls and boys of 0.264 with the 95% confidence interval of (0.078, 0.450), and the MDESD 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 MDESD are affected by the collective effects of those design parameters, some basic conclusions can be drawn: the power will increase (MDESD 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:

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

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In Howard S. Bloom (editor), *Learning More from Social Experiments: Evolving Analytic Approaches*, New York: Russell Sage Foundation.

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

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902. doi:10.1080/19345747.2016.1271069

Borenstein, M. & Hedges, L. V. (2012). CRT-Power – Power Analysis for Cluster-Randomized and Multi-Site Studies [Computer software]. Englewood, NJ: Biostat.

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

Dong, N., Kelcey, B., Spybrook, J. & Maynard, R. A. (2018). PowerUp!-Moderator: A tool for calculating statistical power and minimum detectable effect size differences of the moderator effects in multisite randomized trials. [Software].

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (p. 485). SAGE.

Raudenbush, S. W. & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2), 199-213.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X, & Martinez, A. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 2.01) [Computer software]. www.wtgrantfoundation.org.

Snijders, T. (2001). Sampling. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modeling of health statistics* (pp. 159-173). New York: John Wiley

Snijders, T. (2005). Power and Sample Size in Multilevel Linear Models. In: B.S. Everitt and D.C. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science.* Volume 3, 1570–1573. Chicester(etc.): Wiley, 2005

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics.* Vol. XX, No. X, pp. 1–23. doi: 10.3102/1076998616655442

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the Past Decade: An Examination of the Precision of Cluster Randomized Trials Funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education, 39* (3): 255–267. doi:10.1080/1743727X.2016.1150454.

Weiss, M., Bloom, H. S., & Brock, T. (2014). A Conceptual Framework for Studying the Sources of Variation in Program Effects. *Journal of Policy Analysis and Management, 33* (3), 778–808.

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

1	2	3	4	5	6	7	8	9
Number of Total Levels of Clustering	Model Number	Level of Treatment	Level of Moderator	Slope of Treatment	Binary Moderator		Continuous Moderator	
					MDESD Calculation	Power Calculation	MDESD Calculation	Power Calculation
2	MRT2-1N-1	1	1	Nonrandomly Varying	MRT21N_MDESD	MRT21N_Power	MRT21Nc_MDESD	MRT21Nc_Power
	MRT2-1N-2	1	2	Nonrandomly Varying				
	MRT2-1R-1	1	1	Random	MRT21R_MDESD	MRT21R_Power	MRT21Rc_MDESD	MRT21Rc_Power
	MRT2-1R-2	1	2	Random				
3	MRT3-1N-1	1	1	Nonrandomly Varying	MRT31N_MDESD	MRT31N_Power	MRT31Nc_MDESD	MRT31Nc_Power
	MRT3-1N-2	1	2	Nonrandomly Varying				
	MRT3-1N-3	1	3	Nonrandomly Varying				
	MRT3-1R-1	1	1	Random	MRT31R_MDESD	MRT31R_Power	MRT31Rc_MDESD	MRT31Rc_Power
	MRT3-1R-2	1	2	Random				
	MRT3-1N-3	1	3	Random				
	MRT3-2N-1	2	1	Nonrandomly Varying	MRT32N_MDESD	MRT32N_Power	MRT32Nc_MDESD	MRT32Nc_Power
	MRT3-2N-2	2	2	Nonrandomly Varying				
	MRT3-2N-3	2	3	Nonrandomly Varying				
	MRT3-2R-1	2	1	Random	MRT32R_MDESD	MRT32R_Power	MRT32Rc_MDESD	MRT32Rc_Power
	MRT3-2R-2	2	2	Random				
	MRT3-2R-3	2	3	Random				

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)

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)		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.25	Proportion of variance among sites: $\rho = \tau_{00}^2/(\tau_{00}^2 + \sigma^2)$
ω_3	0.10	The standardized effect variability of the level-1 moderation across sites: $\omega_3 = \tau_{33}^2/\tau_{00}^2$
ω_1	0.30	The standardized effect variability of the treatment effect across sites: $\omega_1 = \tau_{11}^2/\tau_{00}^2$
P	0.40	Proportion of Level 1 units randomized to treatment: $n_T / (n_T + n_C)$
Q_1	0.50	Proportion of Level 1 units in Moderator subgroup: $n_1 / (n_1 + n_0)$
Q_2	0.60	Proportion of Level 2 units in Moderator subgroup: $J_1 / (J_1 + J_0)$
R_1^2	0.50	Proportion of variance in Level 1 outcome explained by Level 1 covariates, moderator, treatment variable, and interaction.
R_2^2	0.10	Proportion of variance for the random treatment slope explained by Level 2 moderator.
n (Average Sample Size for Level 1)	20	Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)
J (Average Sample Size for Level 2)	40	Toal number of Level 2 units (sites)
$MDESD(\lvert\hat{\delta}_{1b}\rvert)$	**0.264**	Minimum Detectable Effect Size Difference regarding Cohen's d for Level-1 Moderator
95% Confidence Interval	**(0.078, 0.45)**	95% Confidence Interval of $MDESD(\lvert\hat{\delta}_{1b}\rvert)$
$MDESD(\lvert\hat{\delta}_{2b}\rvert)$	**0.354**	Minimum Detectable Effect Size Difference regarding Cohen's d for Level-2 Moderator
90% Confidence Interval	**(0.105, 0.603)**	95% Confidence Interval of $MDESD(\lvert\hat{\delta}_{2b}\rvert)$

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
jessaca.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.

Moderator	Study 1 (J = 30)	Study 2 (J = 36)	Study 3 (J = 38)	Study 4 (J = 40)	Study 5 (J = 46)	Study 6 (J = 49)	Study 7 (J = 70)
Urban	0.53	0	0.09	0.30	0.23	0.70	0.86
High Minority*	0.35	1.00	1.00	1.00	0	0.07	0.24
Low SES**	0.39	0.66	0	0.43	0.33	1.00	0.66
Schoolwide Title I	0.69	0.72	0.47	0.48	0.67	0.97	0.79

* 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*.

	School-level ICC	School-level R^2	Individual-level R^2
Student Achievement Outcome	0.15, 0.25	0.50, 0.80	0.40

*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.

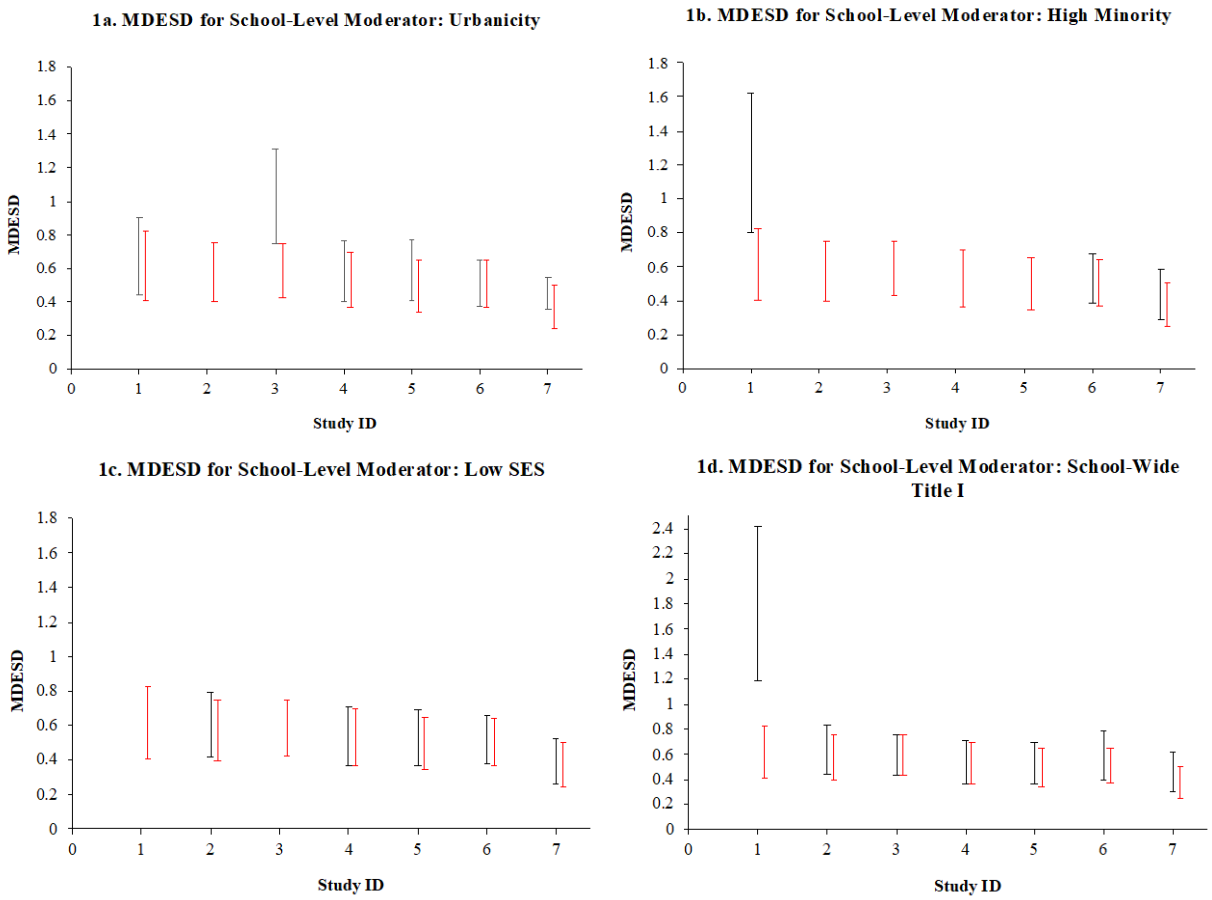


Figure 1. Calculated MDESD based on population proportion (black line) and balanced proportions (red line) for the following moderators: a) Urbanicity, b) High Minority, c) Low SES, d) Schoolwide Title I. Note that MDESD could not be calculated for studies with moderator distribution of either 0 or 1, 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

Dong, N., Kelcey, B., Spybrook, J. & Maynard, R. A. (2016). PowerUp!-Moderator: A tool for calculating statistical power and minimum detectable effect size differences of the moderator effects in cluster randomized trials. [Software]. http://www.causalevaluation.org/.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group-randomized Trials in Education. Educational Evaluation and Policy Analysis, 29(1), 60-87.

Hill, C. J.; Bloom, H. S.; Black, A. R.; & Lipsey, M. (2008). Empirical Benchmark for Interpreting Effect Sizes in Research. *Child Development Perspectives*, 2(3), 172-177.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for Detecting Treatment by Moderator Effect in Two and Three-level Cluster Randomized Trials. *Journal of Educational And Behavior Statistics*, 41(6), 605-627.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the Past Decade: An Examination of the Precision of Cluster Randomized Trials Funded by the U.S. Institute of Education Science. *International Journal of Research & Method in Education*, 39(3), 255-267.

Zhu, P.; Jacob, R.; Bloom, H.; & Xu, Z. (2012). Designing and Analyzing Studies that Randomized Schools to Estimate Intervention Effects on Student Achievement Outcomes without Classroom-level Information. *Educational Evaluation and Policy Analysis*, 34(1), 45-68.

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

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

$$M_{jk} = \pi_{0k} + \pi_1(\bar{X}_{jk} - \bar{X}_k) + \pi_2(W_{jk} - \bar{W}_k) + \pi_3 U_{jk} + \varepsilon_{jk}^M \qquad \varepsilon_{jk}^M \sim N(0, \sigma_{M|}^2) \tag{1}$$

$$\pi_{0k} = \zeta_{00} + aT_k + \zeta_{01}\bar{X}_k + \zeta_{02}\bar{W}_k + \zeta_{03}Z_k + u_{0k}^M \qquad u_{0k}^M \sim N(0, \tau_{M|}^2)$$

Outcome model

$$Y_{ijk} = \beta_{0jk} + \beta_1(X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon_{ijk}^Y \qquad \varepsilon_{ijk}^Y \sim N(0, \sigma_{Y|}^2) \tag{2}$$

$$\beta_{0jk} = \gamma_{00k} + b_{2k}(M_{jk} - \bar{M}_k) + \gamma_{01}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{02}(W_{jk} - \bar{W}_k) + \gamma_{03}U_{jk} u_{0jk}^Y \qquad u_{0jk}^Y \sim N(0, \tau_{Y|}^2)$$

$$\gamma_{00k} = \zeta_0 + B\bar{M}_k + \Delta\bar{M}T_k + c' T_k + \xi_1 \bar{X}_k + \xi_2 \bar{W}_k + \xi_3 Z_k + \upsilon_{00k}^Y \qquad \upsilon_{00k}^Y \sim N(0, \nu_{Y|}^2)$$

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 b_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 υ_{00k}^Y, u_{0jk}^Y and ε_{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 (σ_{aB}) of this product

$$z^{Sobel} = aB / \sqrt{\sigma_{aB}^2} \tag{3}$$

Our omitted derivations extend this test to three-level settings such that the resulting Sobel test for a mediation effect is

$$z_{3L}^{Sobel} = \frac{aB}{\sqrt{\sigma_{aB}^2}} = \frac{aB}{\sqrt{\left(\frac{\tau_M^2(1-R_{M_\omega^{L3}}^2) + (1-R_{M_\omega^{L2}}^2)\sigma_M^2 / n_2}{n_3\sigma_T^2(1-R_T^2)}\right)B^2 + a^2\left(\frac{\nu_Y^2(1-R_{Y^{L3}}^2) + \tau_Y^2(1-R_{Y^{L2}}^2)/n_2 + (1-R_{Y^{L1}}^2)\sigma_Y^2/n_2 n_1}{n_3(\tau_M^2(1-R_{M_\Omega^{L3}}^2) + (1-R_{M_\Omega^{L2}}^2)\sigma_M^2/n_2)}\right)}} \tag{4}$$

We use τ_M^2 and σ_M^2 as the unconditional school- and class-level variances of the mediator, σ_T^2 as variance of the treatment, $R_{M_\omega^{L3}}^2$ and $R_{M_\omega^{L2}}^2$ as the school- and class-level mediator variance explained by predictors in the mediation model, R_T^2 as treatment variance explained by predictors, and n_3 and n_2 as the school- and class-level sample sizes. Similarly we introduce ν_Y^2, τ_Y^2, σ_Y^2, τ_{MT}^2, and σ_{MT}^2 as the unconditional school-, class- and student-level outcome and the unconditional school- and class-level treatment-by-mediator variance and $R_{MT_\Omega^{L3}}^2$, $R_{MT_\Omega^{L2}}^2$, $R_{M_\Omega^{L3}}^2$, and $R_{M_\Omega^{L2}}^2$ 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

$$P(z_{3L}^{Sobel} > z_{critical}) = 1 - \Phi(z_{critical} - z_{3L}^{Sobel}) + \Phi(-z_{critical} - z_{3L}^{Sobel}) \tag{5}$$

where Φ is the normal distribution with $z_{critical}$ 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:

$a = 0.50$ (treatment-mediator relationship [Cohen's d scale])

$B = 0.30$ (mediator-outcome relationship [Standardized regression scale])

$v_Y^2 = 0.10$ (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)

$R_{YL3}^2 = R_{YL2}^2 = R_{YL1}^2 = 0.50$ (outcome variance explained at each level)

$R_{ML3}^2 = R_{ML2}^2 = 0.50$ (mediator variance explained at each level)

$P = 0.50$ (proportion of schools receiving treatment)

$n_2 = 4$ (classrooms/school)

$n_1 = 20$ (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

Desimone, L. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*, 181–199.

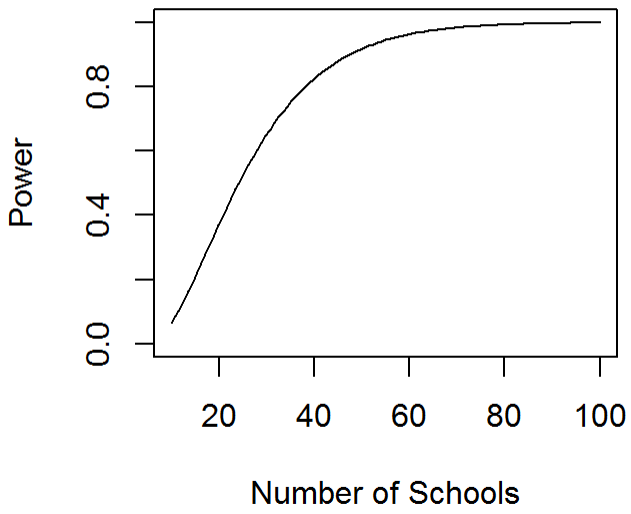
Institute of Education Sciences, U.S. Department of Education & National Science Foundation (2013). *Common Guidelines for Education Research and Development* (NSF 13-126). Retrieved February 15, 2014, from http://ies.ed.gov/pdf/CommonGuidelines.pdf

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

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

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the Past Decade: An Examination of the Precision of Cluster Randomized Trials Funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education, 39, 3, 255-267.*

Figure 1
Power to detect mediation effect in a three-level group-randomized design



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 (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).

$$\begin{aligned} Y_{ij} &= \pi_{0j} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma_Y^2) \\ \pi_{0j} &= \zeta_{00} + \delta T_j + u_{0j} \quad u_{0j} \sim N(0, \tau_{Y|}^2) \end{aligned} . \tag{1}$$

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 π_{0j}, and an individual error term $\varepsilon_{ij} \quad N(0, \sigma_Y^2)$. At the school level, the mean school outcome is decomposed beginning with ζ_{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 δ as it is associated with the treatment indicator T_j.

Measurement error. We consider the use of a fallible measure of the outcome ($\tilde{Y}_{ij}$) rather than the true outcome (Y_{ij}). 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

$$\tilde{Y}_{ij} = Y_{ij} + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e^2) \tag{2}$$

with $\tilde{Y}_{ij}$ as the observed outcome value for student *i* in school *j* and e_{ij} as the independent and normally distributed error with variance σ_e^2. The observed outcome remains normally distributed with individual-level variance σ_Y^2 such that $\sigma_{\tilde{Y}}^2 = \sigma_Y^2 + \sigma_e^2$.

In turn, our analyses probe CRTs with fallibly measured outcomes where the ~ 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

$$\sigma_{\tilde{\delta}}^2 = \frac{\tau_{Y|}^2 + (\sigma_Y^2 / \lambda_Y^{L1}) / n_1}{n_2 \sigma_T^2} \tag{3}$$

with λ_Y^{L1} as the individual-level reliability of the outcome (i.e., $\lambda_Y^{L1} = \frac{\sigma_Y^2}{\sigma_{\tilde{Y}}^2} = \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_e^2}$).

Power. To test inferences regarding the null hypothesis of the treatment effect $H_0 : \tilde{\delta} = 0$ we used the *t*-distribution with

$$t_{\tilde{\delta}} = \frac{\tilde{\delta}}{\sigma_{\tilde{\delta}}} = \frac{\tilde{\delta}}{\sqrt{4(\tau_{Y|}^2 + (\sigma_Y^2 + \sigma_e^2)/n_1)/n_2}} \tag{4}$$

Under the alternative hypothesis $H_1 : \tilde{\delta} > 0$, $t_{\tilde{\delta}}$ follows a non-central *t*-distribution with $t_{\tilde{\delta}}$ as the non-centrality parameter and $df = n_2 - 2$. Power is then determined with

$$P(|t_{\tilde{\delta}}| > t_{n_2-2,\alpha/2}) = 1 - t(t_{n_2-2,\alpha/2} - t_{\tilde{\delta}}) + t(-t_{n_2-2,\alpha/2} - t_{\tilde{\delta}}) \tag{5}$$

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

$$T = c_2 n_2 + c_1 n_2 n_1 \tag{6}$$

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 $\sigma^2_{\bar{\delta}}$ and minimizing this expression in terms of n_1 produces the optimal sample of individuals per cluster when the outcome is measured with (σ^2_e)

$$n_1^{opt^{CRT}} = \sqrt{\left(\frac{c_2}{c_1}\right) \times \left(\frac{\sigma^2_Y + \sigma^2_e}{\tau^2_{Y|}}\right)}. \quad (7)$$

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 $n_1^{opt^{CRT}}$ 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 $n_2 = 280$, $n_1 = 10$, $\sigma^2_Y = .72$, $\tau^2_{Y|} = .28$, and $\delta = 0.2$, we would achieve a power rate of ≈.80. When we include measurement error resulting in $\lambda_Y^{L1} \approx 0.56$, our power rate drops to ≈.74. We could also plan this study using the optimal sample allocation framework to determine n_1 and n_2. With $c_2 / c_1 = 100$, $n_1^{opt^{CRT}} \approx 16$ but when we consider outcome unreliability caused by measurement error $n_1^{opt^{CRT}} \approx 21$.

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

Hull, D. M., Hinerman, K. M., Ferguson, S. L., Chen, Q., & Näslund-Hadley, E. I. (2018). Teacher-led math inquiry: A cluster randomized trial in Belize. *Educational Evaluation and Policy Analysis, 40,* 336-358.

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

Raudenbush, S. W., & Bryk, A. S. (2002*). Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research and Method in Education. 39,* 255–267.

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

σ_e^2	δ	n_2	n_1	σ_Y^2	$\tau_{Y\|}^2$	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					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 (δ) was set at 0.5. δ 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

		Power											
		Delta $\delta = 0.25$						Delta $\delta = 0.1$					
		Sample Size $n_1 : n_2$											
Error (σ_e^2)	Reliability (λ_{L1})	10: 40	20: 40	40: 40	10: 80	20: 80	40: 80	10: 40	20: 40	40: 40	10: 80	20: 80	40: 80
0.00	1.00	0.417	0.521	0.592	0.716	0.826	0.884	0.105	0.123	0.138	0.171	0.209	0.240
0.10	0.90	0.400	0.507	0.583	0.694	0.813	0.878	0.103	0.121	0.136	0.164	0.204	0.236
0.20	0.82	0.383	0.494	0.575	0.673	0.801	0.871	0.100	0.118	0.134	0.159	0.199	0.232
0.30	0.75	0.369	0.481	0.566	0.652	0.788	0.865	0.098	0.116	0.132	0.154	0.194	0.229
0.40	0.69	0.355	0.469	0.558	0.633	0.776	0.858	0.095	0.114	0.130	0.149	0.190	0.225
0.50	0.64	0.342	0.458	0.550	0.614	0.763	0.852	0.094	0.112	0.129	0.145	0.185	0.222
0.60	0.60	0.330	0.447	0.543	0.597	0.751	0.846	0.092	0.110	0.127	0.141	0.181	0.218
0.70	0.56	0.320	0.437	0.535	0.580	0.739	0.839	0.090	0.109	0.126	0.137	0.177	0.215
0.80	0.53	0.309	0.427	0.528	0.564	0.728	0.833	0.089	0.107	0.125	0.134	0.174	0.212
0.90	0.50	0.300	0.417	0.521	0.548	0.716	0.826	0.087	0.105	0.123	0.131	0.171	0.209
% Power loss		28.1	19.8	12.1	23.4	13.3	6.5	17.3	14.6	10.3	23.4	18.6	12.6

Note. In this case $\sigma_Y^2 = 0.9$ and $\tau_{Y|}^2 = 0.1$ 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*

		Cost Ratio: 10					Cost Ratio: 100				
Error (σ_e^2)	Reliability (λ_{L1})	n_1^{opt}	Power	Power Loss	MDE	MDE Loss	n_1^{opt}	Power	Power Loss	MDE	MDE Loss
0.00	1.00	9.487	0.268	0.000	0.509	0.000	30.000	0.549	0.000	0.335	0.000
0.10	0.90	10.000	0.255	0.012	0.523	0.014	31.623	0.538	0.011	0.339	0.004
0.20	0.82	10.488	0.244	0.023	0.537	0.027	33.166	0.528	0.021	0.343	0.008
0.30	0.75	10.954	0.235	0.033	0.550	0.040	34.641	0.518	0.030	0.347	0.012
0.40	0.69	11.402	0.226	0.042	0.562	0.053	36.056	0.510	0.039	0.351	0.016
0.50	0.64	11.832	0.218	0.050	0.574	0.065	37.417	0.501	0.047	0.354	0.020
0.60	0.60	12.247	0.211	0.057	0.585	0.076	38.730	0.493	0.055	0.358	0.023
0.70	0.56	12.649	0.204	0.064	0.597	0.087	40.000	0.486	0.063	0.361	0.027
0.80	0.53	13.038	0.198	0.070	0.608	0.098	41.231	0.479	0.070	0.365	0.030
0.90	0.50	13.416	0.192	0.076	0.618	0.109	42.426	0.472	0.077	0.368	0.033
% Change		41.4	-28.2		21.3		41.4	-14.0		9.9	

Note. In this selected condition $\delta = 0.25$ and $\sigma_Y^2 = 0.9$ $\tau_{Y|}^2 = 0.1$.

FIGURE 1

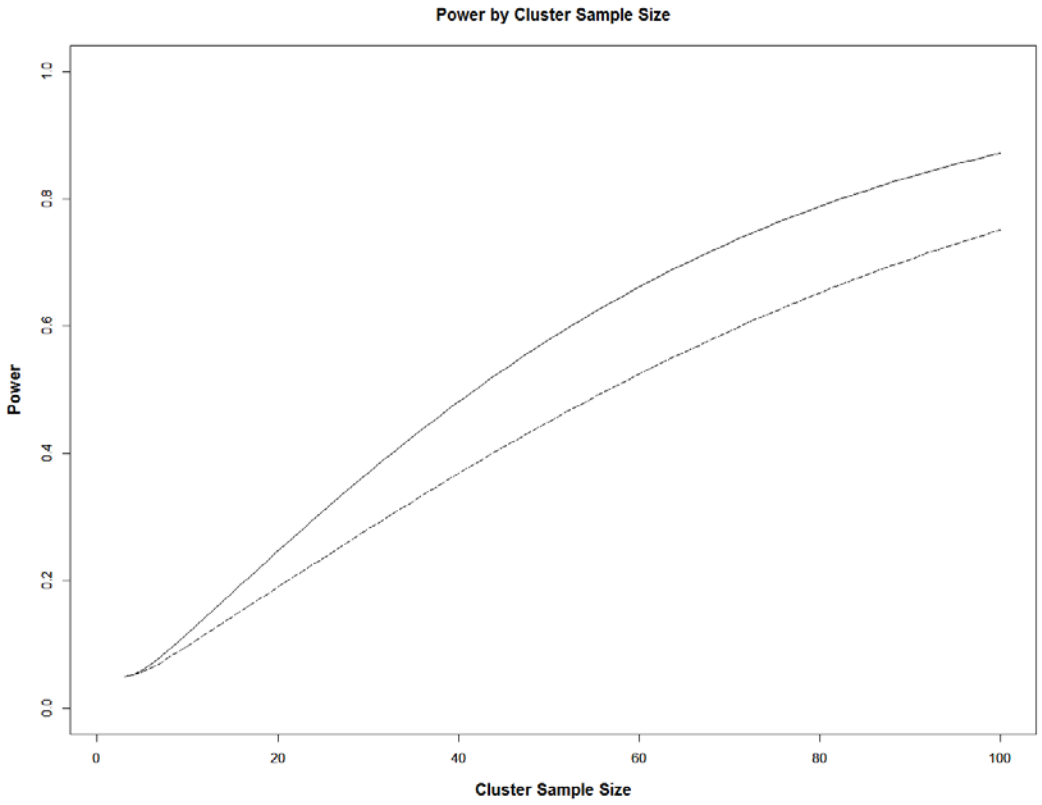


FIGURE 1. Power to detect the main effect in a two-level cluster randomized study as a function of cluster sample size when individual- to cluster-variance in the outcome is 0.9:0.1, the magnitude of the treatment effect is 0.25, the individual per cluster sample size is 10, and there is no measurement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

FIGURE 2

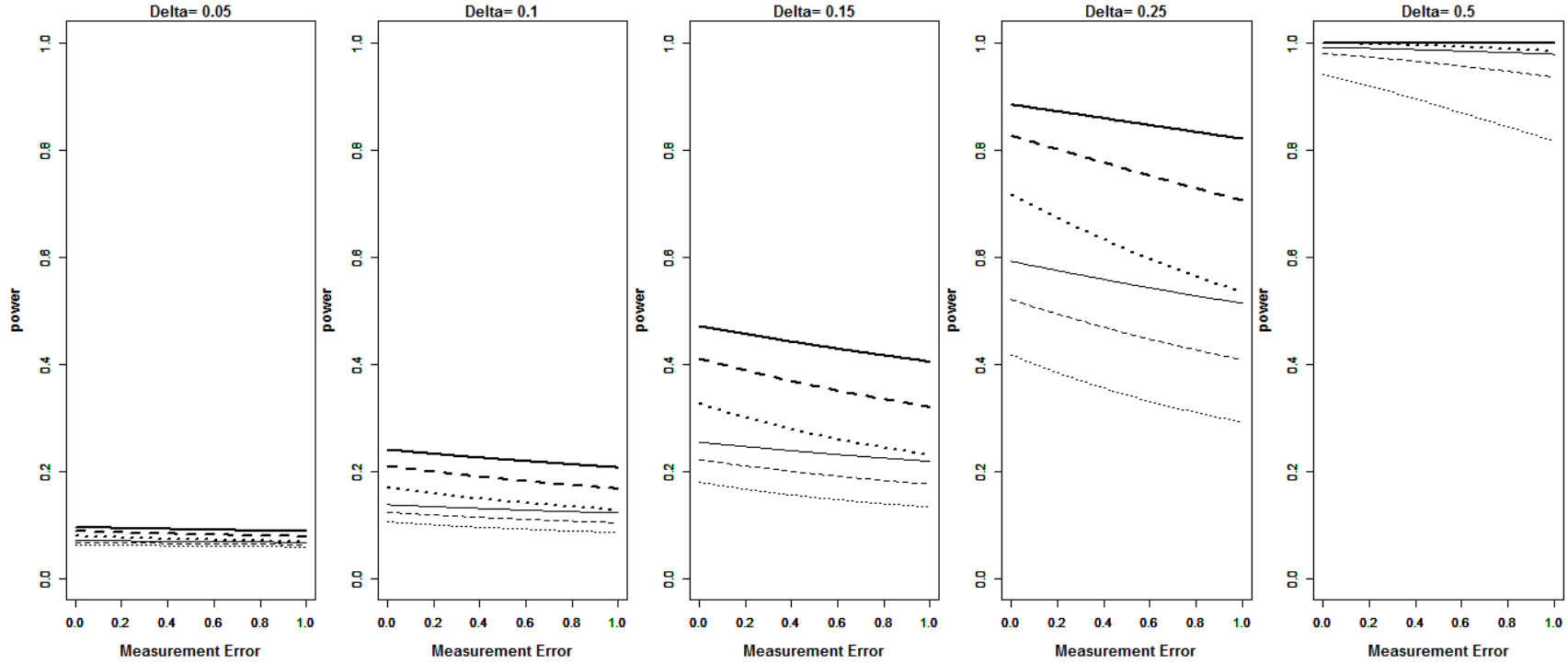


FIGURE 1. Power to detect a main effect in a two-level cluster randomized study as a function of measurement error in the level-one outcome when individual to cluster variance is 0.9:0.1, delta is 0.05, 0.1, 0.15, 0.25, and 0.5 and the individual sample size is 10 (dot), 20 (dash), and 40 (solid) with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

FIGURE 3

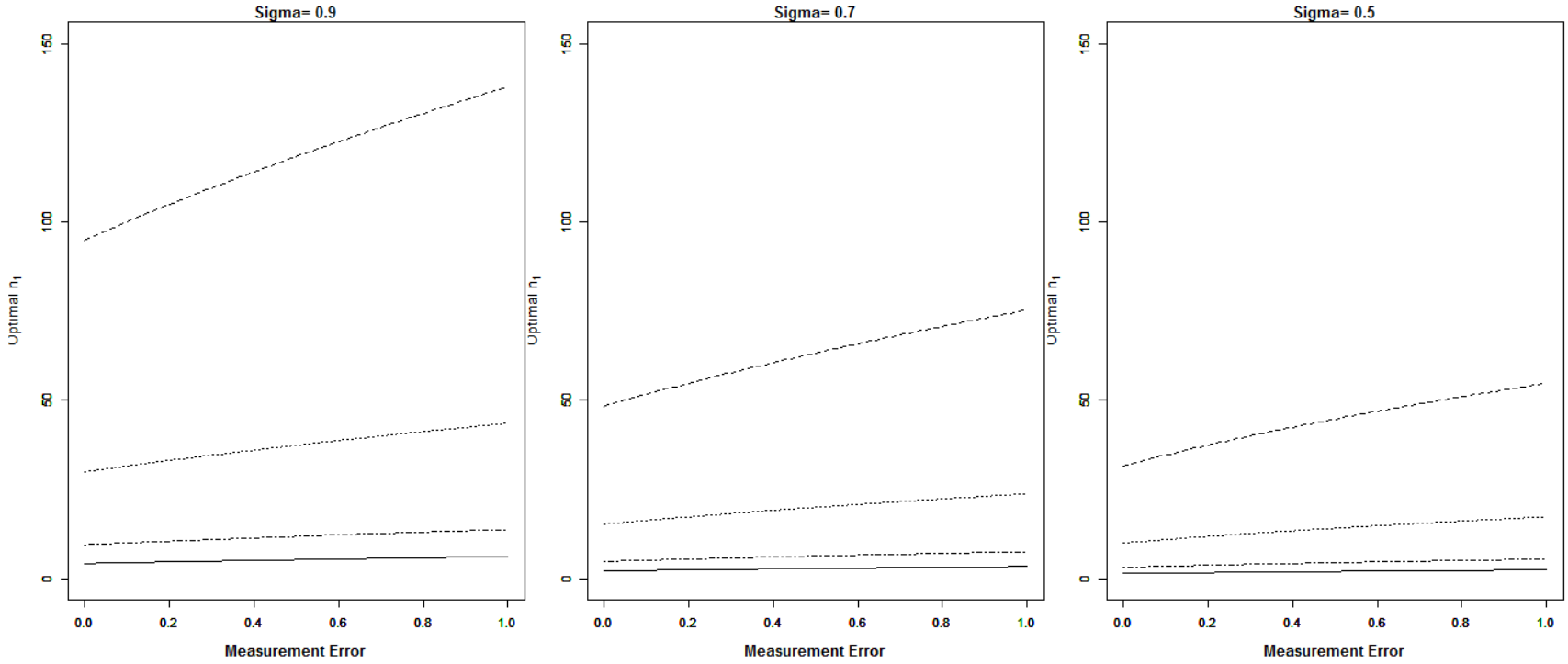


FIGURE 3. Optimal individual sample size for two-level cluster randomized studies as a function of measurement error in the level-one outcome when individual to cluster level variance is 0.9:0.1; 0.7:0.3; and 0.5:0.5 and the group to individual cost ratio is 1000:1 (dashed), 100:1 (dot). 10:1 (dash/dot), and 2:1 (solid).