Dong, N., Spybrook, J., Kelcey, B., & Bulus, M. (in press). Power analyses for moderator effects with (non)random slopes in cluster randomized trials. *Methodology*.

Authors and Affiliations:

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

Jessaca Spybrook
Western Michigan University
3571 Sangren Hall
Kalamazoo, Michigan 49008
Phone: (269) 387-3889
jessaca.spybrook@wmich.edu

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

Metin Bulus
Adiyaman University
Adıyaman, Turkey
bulusmetin@gmail.com

*Corresponding Author

This project has been funded by the National Science Foundation [1437679, 1437692, 1437745, 1913563]. The opinions expressed herein are those of the authors and not the funding agency.

Power Analyses for Moderator Effects with (Non)Randomly Varying Slopes in Cluster Randomized Trials

Abstract

Researchers often apply moderation analyses to examine whether the effects of an intervention differ conditional on individual or cluster moderator variables such as gender, pretest, or school size. This study develops formulas for power analyses to detect moderator effects in two-level cluster randomized trials (CRTs) using linear models. We derive the formulas for estimating statistical power, minimum detectable effect size difference and 95% confidence intervals for cluster- and individual-level moderators. Our framework accommodates binary or continuous moderators, designs with or without covariates, and effects of individual-level moderators that vary randomly or nonrandomly across clusters. A small Monte Carlo simulation confirms the accuracy of our formulas. We also compare power between main effect analysis and moderation analysis, discuss the effects of mis-specification of the moderator slope (randomly vs. non-randomly varying), and conclude with directions for future research. We provide software for conducting a power analysis of moderator effects in CRTs.

Key words: *cluster randomized trials (CRTs), minimum detectable effect size difference, moderator effect, statistical power*

Power Analyses for Moderator Effects with (Non)Randomly Varying Slopes in Cluster Randomized Trials

A critical consideration in the evaluation of treatment programs is whether those treatment effects are moderated by context or individual characteristics. As a result, an important consideration that emerges in the planning stage is how to design studies that have the sufficient power to detect such moderation if it exists. Although there has been a steady pace of advancement in the design of moderation studies in cluster randomized trials (CRTs; Bloom, 2005; Dong, Spybrook, & Kelcey, 2018; Mathieu, Aguinis, Culpepper, & Chen, 2012; Moerbeek & Maas, 2005, Spybrook, Kelcey, & Dong, 2016), extant studies are largely fragmented in that they normally consider only isolated aspects of the design rather than the full assembly of design considerations that are typically encountered in planning such a study. For instance, with the exception of a few studies (e.g., Dong, Spybrook, & Kelcey, 2018), prior literature regarding the estimation of statistical power for moderation has often limited its analysis to only binary moderators or has failed to include additional covariates (i.e., "unconditional designs"; Bloom, 2005; Spybrook, Kelcey, & Dong, 2016). Given the widespread presence of moderators that are continuous in nature (e.g., pretest) and the widespread use of covariate-adjusted designs to improve power and reduce potential bias due to unhappy randomization, it is critical to provide a more general set of tools for power analyses that can readily accommodate such variations (e.g., Bloom, 2006; Bloom, Richburg-Hayes & Rebeck Black, 2007; Dong & Maynard, 2013; Moerbeek, 2006; Moerbeek, van Breukelen, & Berger, 2001; Raudenbush, Martinez, & Spybrook, 2007).

Similarly, current multilevel literature is limited in the guidance it offers concerning statistical power when assessing the extent to which treatment effects vary across subgroups defined by an individual-level variable. More specifically, assessments of individual-level moderators are typically operationalized through cross-level interactions between the cluster-level treatments and individual-level moderators (e.g., child's gender). The result is that the effect of the individual-level variable (i.e., as quantified through the coefficient) can be regarded as randomly or nonrandomly varying across clusters. The nonrandomly varying slope approach assumes that the gender achievement gap does not vary randomly across schools but rather only as an explicit function of cluster-level variables (e.g., the individual-level slope or coefficient for gender varies across clusters only as a function of the treatment status). The randomly varying slope or coefficient model addresses the same moderation question, but allows for the possibility that the gender achievement slope or coefficient randomly varies across schools even after accounting for the treatment effect (e.g., unexplained heterogeneity across schools in terms of the relationship between gender and the outcome). The choice between these approaches ultimately depends on prior knowledge of the effects of the moderator variables and the theory underlying the intervention. However, it is important that design frameworks consider both of these approaches and the implications of designing a study based on one of the frameworks.

Our review of the literature identified only two methodological studies that have examined the power for the randomly varying slope model in moderation analysis (Dong, Spybrook, & Kelcey, 2018; Mathieu, Aguinis, Culpepper, & Chen, 2012). In addition, there are no studies that have examined the trade-offs between the design assumptions, the effects on power when the slope is mis-specified (randomly vs. non-randomly varying slope) or the potential inaccuracies that accumulate in power formulas under such mis-specifications. A mis-specification of the slope term potentially undermines the accuracy of the standard error estimates for the moderator effect, which may result in incorrect estimates of statistical power. Investigation of the effects of a mis-specified slope can help us understand how much the bias on power arises due to either type of mis-specification, helps develop potential strategies to mitigate bias due to such mis-specifications, and ultimately to design moderation studies that are robust and well-positioned to detect such effects.

A key prior contribution to the literature with regard to designing multilevel moderation studies was Mathieu et al. (2012). Mathieu et al. (2012) conducted a comprehensive Monte Carlo simulation to estimate the statistical power to detect cross-level interaction effects in multilevel modeling. However, Mathieu et al (2012) only studied two-level models without including covariate adjustment on additional covariates separate from the moderator, and did not provide closed form formulas to estimate the statistical power, minimum detectable effect size difference (MDESD) between moderator subgroups, or minimum required sample size to detect meaningful effects. Dong, Spybrook, and Kelcey (2018) extended this line of inquiry by developing the formulas to calculate statistical power and MDESD by considering the levels of the moderators at which they have been assessed, the distribution of moderators (binary vs. continuous), the slopes of lower level moderators (random vs. non-randomly varying), and the level of covariates for three-level CRTs. However, the scope, developments and analyses in Dong, Spybrook, and Kelcey (2018) did not cover two-level CRTs.

The purpose of this study is to consolidate and extend the literature on power analyses for moderators by developing power formulas that accommodate categorical or continuous moderators, models with or without covariates, same or cross-level moderator effects, and nonrandomly varying or randomly varying slopes in two-level CRTs. We then advance the practical application of these results by examining the effects on power when the slope is mis-specified (randomly varying slope vs. non-randomly varying slope) to outline the sensitivity of power analysis to such mis-specifications. Because a team planning a CRT may be interested in the power for a moderator effect of a given magnitude or the MDESD given sample size and the desired power, we provide the power formulas as well as the MDESD calculations and their corresponding confidence intervals. We also created a Microsoft Excel-based function, an R function, and an R shinny app to assist researchers conducting power analyses for various moderator effects[1].

The paper is organized as follows. We present the formulas for statistical power and the MDESD and its confidence intervals for the moderator variable at level 2 and subsequently for a moderator at level 1. In each case, we start with a continuous moderator and extend it to a binary moderator. We also conduct a small Monte Carlo simulation to assess the empirical validity of the formulas in finite sample sizes. We then compare the statistical power and MDESD for moderation effects under different design considerations followed by a comparison of the MDES for main treatment effects and the MDESD for the moderation effects. Finally, we discuss the implications of planning studies to detect moderator effects in two-level CRTs and consider directions for future work.

Statistical Power and Minimum Detectable Effect Size Difference in Two-Level CRTs

We present the key results of the formulas for statistical power and the MDESD and its confidence intervals for different moderator effects in the framework of a two-level hierarchical linear model (HLM; Raudenbush & Bryk, 2002). The detailed derivations are in Electronic Supplementary Material 1.

Two-level CRTs with a Moderator at Level 2

[1] The software can be accessed from the website: https://www.causalevaluation.org/.

We begin with a two-level design that randomly assigns groups/clusters (e.g., schools) to the treatment or control condition and conditions on a cluster-level covariate (e.g., the percentage of students eligible for free or reduced-price lunch) and probes a cluster-level moderator (e.g., school size). The data are generated using a two-level hierarchical linear model (Raudenbush & Bryk, 2002):

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij}, r_{ij} \sim N(0, \sigma^2_{|X}) \quad (1)$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}S_j + \gamma_{02}T_j + \gamma_{03}(S_j \times T_j) + \gamma_{04}W_j + \gamma_{05}\bar{X}_{.j} + u_{0j},\ u_{0j} \sim N(0, \tau^2_{|S,W,\bar{X},T}).$$
$$\beta_{1j} = \gamma_{10}. \quad (2)$$

Y_{ij} is the outcome measure for observation i (i = 1,…,n_j) in cluster j (j = 1,…, J), T_j is a binary variable indicating the treatment status coded as ±½, S_j is a level-2 continuous moderator ($S_j \sim N(0, S_s^2)$), X_{ij} is a level-1 covariate and $\bar{X}_{.j}$ is the sample group mean, and W_j is a level-2 covariate ($W_j \sim N(0, S_w^2)$). r_{ij} is the level-1 random error, $r_{ij} \sim N(0, \sigma^2_{|X})$, and u_{0j} is the random effect for the intercepts, $u_{0j} \sim N(0, \tau^2_{|S,W,\bar{X},T})$. As in the single level regression analysis, centering variables yields desirable statistical properties (Aiken & West, 1991), group-mean centering is used in Equation 1 to gain some computational and derivational advantages. Note that in random intercept models, parameter estimates under group-mean centering, grand-mean centering, and no centering can be equated using simple transformations (e.g., Kreft, de Leeuw, & Aiken, 1995). γ_{02} and γ_{03} represent the main effect of treatment and moderator effect, respectively.

We assume that the data are balanced such that each cluster has the same number of observations (n_j= n). However, we do not assume the clusters are equally allocated to treatment conditions. Although equal allocation of clusters to the treatment and control conditions typically yields the most sensitive design (i.e., highest power to detect main and moderator effects), such balance is not always possible in reality. For this reason, we considered a more flexible approach that introduces *P* as the proportion of total clusters that are randomly assigned to the treatment group.

We can test γ_{03} using a *t*-test. Assuming the alternative hypothesis is true, the test statistic follows a non-central *t*-distribution, *T'*, and the standardized noncentrality parameter is:

$$\lambda_{|S,W,X} = \sqrt{\frac{\delta_{2c}^2 P(1-P)(J-6)}{(1-R_2^2)\rho+(1-R_1^2)(1-\rho)/n}}, \quad (3)$$

where *J* is the number of total clusters, *n* is sample size for every cluster (e.g., number of students per school), *P* is the proportion of total clusters that are randomly assigned to the treatment group. R_2^2 is the proportion of variance at level 2 that is explained by the level-2 predictors (S_j ,W_j, T_j, $\bar{X}_{\cdot j}$, and $(S_j \times T_j)$): $R_2^2 = 1 - \frac{\tau_{|S,W,\bar{X},T}^2}{\tau^2}$, where τ^2 is the unconditional level-2 variance; R_1^2 is the proportion of variance at level 1 that is explained by the level-1 predictor $(X_{ij} - \bar{X}_{.j})$, $R_1^2 = 1 - \frac{\sigma_{|X}^2}{\sigma^2}$, where σ^2 is the unconditional level-1 variance. ρ is the unconditional intraclass correlation, $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$. δ_{2c} is the standardized coefficient of $(S_j \times T_j)$, (where the subscript indicates the use of a level-2 continuous moderator) such that $\delta_{2c} = \hat{\gamma}_{03}\sqrt{\frac{S_S^2}{\tau^2 + \sigma^2}}$, where S_S^2 is the variance of S_j.

The statistical power for a two-sided test is (note t_0=$t_{1-\alpha/2,J-6}$):

$1-\beta = 1 - P\left[T'\left(J-6, \lambda_{|S,W,X}\right) < t_0\right] + P\left[T'\left(J-6, \lambda_{|S,W,X}\right) \leq -t_0\right]$, where the degrees of freedom[2] is $v = J-6$.

The MDESD for the standardized coefficient is:

$$MDESD(|\delta_{2c}|) = M_v \sqrt{\frac{\left(1-R_2^2\right)\rho + \left(1-R_1^2\right)(1-\rho)/n}{P(1-P)(J-6)}}, \tag{4}$$

where, $M_v = t_\alpha + t_{1-\beta}$ for one-tailed tests with v degrees of freedom ($v = J-6$), and $M_v = t_{\alpha/2} + t_{1-\beta}$ for two-tailed tests.

The 100*(1-α)% confidence interval for $MDESD(|\delta_{2c}|)$ is given by:

$$(M_v \pm t_{\alpha/2} \sqrt{\frac{\left(1-R_2^2\right)\rho + (1-\rho)\left(1-R_1^2\right)/n}{P(1-P)(J-6)}}. \tag{5}$$

When the moderator, S_j, is a binary variable with a proportion of Q in one moderator subgroup and (1-Q) in another moderator subgroup, the standardized noncentrality parameter is:

$$\lambda_{|S,W} = \sqrt{\frac{\delta_{2b}^2 P(1-P)Q(1-Q)(J-5)}{\left(1-R_2^2\right)\rho + (1-\rho)/n}}, \tag{6}$$

where δ_{2b} is the effect size (standardized mean difference), $\delta_{2b} = \hat{\gamma}_{03} / \sqrt{\tau^2 + \sigma^2}$.

Table 1 presents the summary of standardized noncentrality parameters, MDESD and 100*(1-α)% confidence intervals, and degrees of freedom for the t-test for various two-level moderation models. The above results are presented under Model "CRT2-2", which stands for a two-level CRT with a level-2 moderator and flexible treatment allocation. Note that we assume the fixed slope for covariate $(X_{ij} - \bar{X}_{.j})$ in Equation 2 for the purpose of simplicity. Because the moderation term is in the equation for the level-2 intercept, the standard error of the moderator

[2] Generally, $v = J - g^* - 4$, where g^* is the number of Level 2 covariates (excluding the treatment variable, moderator, and moderator*treatment).

effect is not affected by the slopes of other level-1 covariates, hence, the power and MDESD formulas apply to the model with random slope for $(X_{ij} - \bar{X}_{.j})$.

[Table 1 about here]

Two-level CRTs with a Moderator at Level 1

Under the same design, we next consider individual-level moderators allowing for two different specifications: (1) the randomly varying slope model, which assumes that the effect of the level-1 moderator varies by the treatment status and varies randomly across the level-2 units, and (2) the nonrandomly varying slope model, which assumes that the effect of the level-1 moderator varies by the treatment status but does not vary further across the level-2 units.

The Randomly Varying Slope Model. The randomly varying slope hierarchical linear model, including one treatment variable, T_j, and one level-1 moderator, S_{ij} ($S_{ij} \sim N(0, S_s^2)$), with a random slope is:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} S_{ij} + r_{ij},\ r_{ij} \sim N(0, \sigma_{|S}^2). \tag{7}$$

Level 2:

$$\begin{array}{l} \beta_{0j} = \gamma_{00} + \gamma_{01} T_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} T_j + u_{1j} \end{array},\ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00|T}^2 & \tau_{01|T} \\ \tau_{10|T} & \tau_{11|T}^2 \end{pmatrix} \right). \tag{8}$$

The level-2 residuals for the intercept, u_{0j}, and the slope, u_{1j}, conditional on the treatment status, have a multivariate normal distribution with means of 0. $\tau_{00|T}^2$ and $\tau_{11|T}^2$ are the variances, and $\tau_{01|T}$ is the covariance for u_{0j} and u_{1j} conditional on the treatment status. The parameter of interest for the moderator effect is γ_{11}. Note that in the context of CRTs, we treat the treatment status (T_j) as the focal predictor and S_{ij} as the moderator, and interpret γ_{11} as the treatment effect of T_j depending on S_{ij}. We may also interpret γ_{11} as the effect of S_{ij} on the outcome depending on the treatment status (T_j).

We test the moderator effect (γ_{11}) using a *t*-test. Based on the formula for the variance of the estimated regression coefficients of a level-1 variable with random slope (Snijders, 2001, 2005), we can derive the standardized noncentrality parameter as below:

$$\lambda_{|S} = \sqrt{\frac{\delta_{1c}^2 P(1-P)J}{(1-R_{2T}^2)\rho\omega + (1-R_1^2)(1-\rho)/n}} \,. \tag{9}$$

ρ is the unconditional intraclass correlation, $\rho = \dfrac{\tau_{00}^2}{\tau_{00}^2 + \sigma^2}$, where σ^2 and τ_{00}^2 are the variances of residuals for level-1 and level-2 intercept in the unconditional model without any predictors. R_1^2 is the proportion of variance at level 1 that is explained by the level-1 moderator (S_{ij}): $R_1^2 = 1 - \dfrac{\sigma_{|S}^2}{\sigma^2}$. R_{2T}^2 is the proportion of the random slope (for *S*) variance explained by the treatment indicator (T_j): $R_{2T}^2 = 1 - \dfrac{\tau_{11|T}^2}{\tau_{11}^2}$. ω is the proportion of the variance (τ_{11}^2) between clusters on the effect of S_{ij} to the between-cluster residual variance (τ_{00}^2) when $\tau_{00}^2 > 0$ under the multilevel modeling framework, $\omega = \dfrac{\tau_{11}^2}{\tau_{00}^2}$. ω indicates the effect heterogeneity for the level-1 moderator (S_{ij}) across level-2 units (clusters) in the model that is not conditional on the treatment variable, T_j. *P* is the proportion of clusters in the treatment group. δ_{1c} is the standardized coefficient, $\delta_{1c} = \hat{\gamma}_{11}\sqrt{\frac{S_S^2}{\tau_{00}^2 + \sigma^2}}$, where S_S^2 is the variance of S_{ij}.

The statistical power for a two-sided test is (note $t_0 = t_{1-\alpha/2, J-2}$):

$1-\beta = 1 - P[T^{'}(J-2, \lambda_{|S}) < t_0] + P[T^{'}(J-2, \lambda_{|S}) \leq -t_0]$, where the degrees of freedom is $v = J - 2$.

The MDESD for the standardized coefficient is:

$$MDESD(|\delta_{1c}|) = M_v \sqrt{\frac{(1-R_{2T}^2)\rho\omega + (1-R_1^2)(1-\rho)/n}{P(1-P)J}}, \tag{10}$$

where, $M_v = t_\alpha + t_{1-\beta}$ for one-tailed tests with v degrees of freedom ($v = J - 2$), and

$M_v = t_{\alpha/2} + t_{1-\beta}$ for two-tailed tests.

The 100*(1-α)% confidence interval for $MDESD(|\delta_{1c}|)$ is given by:

$$(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R_{2T}^2)\rho\omega + (1-R_1^2)(1-\rho)/n}{P(1-P)J}}. \tag{11}$$

The Nonrandomly Varying Slope Model. In the nonrandomly varying slope model the Level 1 model is the same as that in equation (7). However, the Level 2 model is:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}T_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}T_j \end{aligned}, \quad u_{0j} \sim N(0, \tau_{|T}^2). \tag{12}$$

The standardized noncentrality parameter is:

$$\lambda_{|S} = \sqrt{\frac{\delta_{1c}^2 P(1-P)Jn}{(1-R_1^2)(1-\rho)}}. \tag{13}$$

The degrees of freedom[3] is $v = J(n-1) - 2$.

Extension to Binary Moderator. When the level-1 moderator, S_{ij}, is a binary variable with

[3] Generally, $v = J(n-1) - 2 - g^*$, where g^* is the number of Level 1 covariates (excluding the moderator).

a proportion of *Q* in one moderator subgroup and (1 - *Q*) in another moderator subgroup, the noncentrality parameters (standardized) for the randomly varying slope model and the nonrandomly varying slope model are:

$$\lambda_{|S} = \sqrt{\frac{\delta_{1b}^2 P(1-P)J}{(1-R_{2T}^2)\rho\omega + (1-R_1^2)(1-\rho)/(nQ(1-Q))}}\ , \tag{14}$$

and

$$\lambda_{|S} = \sqrt{\frac{\delta_{1b}^2 P(1-P)Q(1-Q)Jn}{(1-R_1^2)(1-\rho)}}\ , \tag{15}$$

where δ_{1b} is the effect size (standardized mean difference), $\delta_{1b} = \hat{\gamma}_{11} / \sqrt{\tau_{00}^2 + \sigma^2}$.

The standardized noncentrality parameters, the minimum detectable effect size difference (MDESD) for the standardized regression coefficient, and the 100*(1-*α*)% confidence interval for $MDESD(|\delta_{1c}|)$ for a continuous level-1 moderator with randomly varying slope and nonrandomly varying slope are presented under Models “CRT2-1R” and “CRT2-1N” in Table 1. The MDESD for the standardized mean difference, and the 100*(1-*α*)% confidence interval for $MDESD(|\delta_{1b}|)$ for a binary level-1 moderator with randomly varying slope and nonrandomly varying slope are presented under Models “CRT2-1R” and “CRT2-1N” in Table 1.

Monte Carlo Simulation

To validate the standard error and power formulas we derived, we conducted a small Monte Carlo simulation. The simulation results provided initial but limited evidence of the close correspondence on the standard error and power (or Type I error) between our formulas and the empirical distribution from the simulation when the analytic model was correctly specified. The detailed procedures and results are presented in Electronic Supplementary Material 2.

We note one particular finding that emerges from the results of the simulation. For a level-1 moderator, we set the effect heterogeneity (ω) for the level-1 moderator across level-2 units varied from 0 to 0.8. For each dataset, we used both the randomly varying slope model and the nonrandomly varying slope model to estimate the moderator effects. When ω is set as 0, the nonrandomly varying slope model is the correctly specified analytic model while the randomly varying slope model is mis-specified analytic model. In these simulations, the randomly varying slope model tended to slightly over-estimate the standard error, but the coverage rate of 95% CI is as good as the nonrandomly varying slope model. Comparing with the nonrandomly varying slope model, the randomly varying slope model produced slightly smaller power. When ω is set as 0.2, 0.4, 0.6, and 0.8, the nonrandomly varying slope model is the mis-specified analytic model while the randomly varying slope model is the correctly specified analytic model (See Tables 9-24 in Electronic Supplementary Material 2). In these simulations, the randomly varying slope model produced closer estimates of the standard error and the coverage rate of 95% CI than the nonrandomly varying slope model. The nonrandomly varying slope model produced bigger bias in the standard error estimates and worse coverage rage of 95% CI when ω increases. Bias in the standard error estimates for mis-specified models are consistent with LaHuis et al's (in press) findings. Figure 1 below clearly demonstrates the relationship between the standard error (SE) and the coverage rate of 95% CI with the heterogeneity coefficient (ω).

[Figure 1 about here]

Discussion: Comparisons among Moderation Designs and Main Effect Designs

Contrasting Moderation Designs

As in the power analysis of the main treatment effect, the power of the moderator effect in two-level CRTs is associated with the noncentrality parameter (λ) and the critical t value (t_0).

The critical *t* value (t_0) is associated with the degrees of freedom (v), the Type I error rate (α), and the choice of a one-tailed or two-tailed test. The noncentrality parameter (λ) is a ratio of the moderator effect estimate to its standard error (SE), which is a function of the total number of clusters (J) and the number of individuals per cluster (n), the proportion of clusters in the treatment group (P), the proportion of variance at level-2 explained by covariates (R_2^2), and the unconditional intraclass correlation (ICC).

If the moderator is a binary variable, the power is also associated with the proportion (Q) of the sample in one moderator subgroup. The MDESD using the standardized mean difference for the binary moderators is $\sqrt{Q(1-Q)}$ times larger than the MDESD using the standardized regression coefficient for the continuous moderators when the moderators are at level 2 or level 1 with the nonrandomly varying slopes. When the sample is equally allocated between the moderator subgroups (Q = 0.5), the design has the biggest power (smallest MDESD) among all options of Q that ranges from 0 to 1.

If the moderator is at level-1 with a randomly varying slope, the power is also associated with the effect heterogeneity (ω) for the level-1 moderator across level-2 units. The MDESD increases and power decreases as ω increases. The results for the nonrandomly varying slope model for the level-1 moderator do not contain the factor that is related to ω. The degrees of freedom also differ depending on whether it is a random slope model or not. The degree of freedom (v) is $J(n-1)-2$ for the nonrandomly varying model while $v=J-2$ for the randomly varying slope model. This is because the interaction term of the treatment and moderator variables varies among the level-1 units within each level-2 cluster for the nonrandomly varying model, but the level-2 random term (i.e., u_{1j} in Expression 8) associated with the coefficient of the moderator in the randomly varying slope model varies among the level-2 clusters. As a result, when the estimation models are correctly specified for the real data, the model with a varying moderator slope will yield less precise estimates than the model with a constant moderator slope. The differences for the power and MDESD between the two models decreases when the number of clusters (*J*) increases and the effect heterogeneity (ω) decreases.

Using the mis-specified analytic models for study design will result in either overestimating or underestimating the power. Specifically, if the randomly varying slope model is used to design the studies where $\omega = 0$, the power will be underestimated; if the nonrandomly varying model is used to design the studies where $\omega > 0$, the power will be overestimated. The bias in power estimates due to model mis-specification decrease when the sample size for the clusters (*J*) increases and the effect heterogeneity (ω) decreases.

To make these comparisons more concrete, we compare MDESD and power among three moderation designs using several examples. Suppose a team of researchers are designing a two-level CRT to test the efficacy of a school-based intervention on student achievement. They are interested in student-level moderator effects and school-level moderator effects. They approach the moderator power analyses from two perspectives: (1) what is the MDESD given power of 0.80 and (2) what is the power for a moderation effect size of 0.20. Based on the literature (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007, 2014) they assume an intraclass correlation coefficient (ρ) of 0.23, and the proportions of variance explained by the covariates at level 1 and level 2 of 0.5 ($R_1^2 = R_2^2$ =0.5). To be conservative, they assume the proportion of variance between schools on the effect of the student-level moderator explained by the school-level predictor to be 0 (R_{2T}^2 = 0). The effect heterogeneity (ω) for the student-level moderator across school-levels is assumed as 0.3 for the randomly varying slope model, which is equivalent to an effect size variability of 0.069 (= $\rho \times \omega$ = 0.23 × 0.3). They use a balanced design with equal assignment of schools to the treatment and control groups (P = 0.5) and 100 students per school. They are interested in the results for a binary moderator and a continuous moderator. For the binary case, they assume half of the sample is in one moderator subgroup (Q = 0.5). Table 2 shows the results of MDESD and power for the total numbers (J) of schools of 40 and 80 under the above assumptions.

[Table 2 about here]

The findings in Table 2 are discussed below. First, a design always has a smaller MDESD, or larger power for a fixed effect size when the level-2 sample size is bigger. Second, the MDESD is larger or the power is smaller for a fixed effect size when the moderator is at the school level compared to the student level. Third, when the moderator is at the student level, the nonrandomly varying moderator slope model has a smaller MDESD, or bigger power for a fixed effect size than the random moderator slope model. Finally, the MDESD as defined by the standardized mean difference for the binary moderator and Q = 0.5 is always twice the value of the MDESD defined by the standardized coefficient for the continuous moderator when the moderator is at the school level or the moderator is at the student level with the nonrandomly varying slope.

Comparing Moderation Designs with Main Effect Designs

We examine the ratio of the MDESD for the moderator analysis to the minimum detectable effect size (MDES) for the main effect analysis. The MDES formula for a two-level cluster randomized design with a level-1 and two level-2 covariates is as follows (Bloom, 2006):

$$MDES = M_{J-4}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)Jn}}\ , \qquad (16)$$

where the multiplier $M_{J\text{-}4} = t_{\alpha/2} + t_{1-\beta}$ with *J*- 4 degrees of freedom.

We use the MDESD formulas for binary moderators in Table 1. The ratio of MDESD for a level-2 binary moderator to the MDES of the main effect when there is no level-1 covariate is:

$$\frac{MDESD_{CRT2-2}}{MDES} = \frac{M_{J-6}}{M_{J-4}}\sqrt{\frac{J-6}{JQ(1-Q)}}. \tag{17}$$

The result in Expression 17 is consistent with Bloom (2005) except Expression 17 includes an extra factor $\sqrt{\frac{J-6}{J}}$. Bloom (2005) derived the standard error of the moderator effects based on the population using the sample size *J* while we derived the standard error based on the sample by adjusting for the degrees of freedom using *J* - 6 (our Monte Carlo simulation suggested that our formulas worked better especially when the sample size is small). $\frac{MDESD_{CRT2-2}}{MDES}$ is around 2 when it is a balanced design (*Q* = 0.5) and there is a large sample size ($\frac{M_{J-6}}{M_{J-4}}$ is close to 1 when *J* is larger than 10, e.g., $\frac{M_{J-6}}{M_{J-4}}$ = 1.01 when *J* = 11.). This result indicates that the MDESD for a level-2 moderator is about twice as large as the MDES of the main effect using the same set of covariates in both cases in the same study. This is analogous to using the ordinary least square (OLS) regression to analyze the completely randomized trials, which do not involve hierarchical data. This makes the level-2 moderator effect more difficult to detect than the main effect just as in the OLS analysis of the completely randomized trials.

The situation is different for the analysis of the level-1 moderator effect, which may have bigger power than the main effect. The MDES formula for the main effect in Expression 16 includes an additional component that is associated with the level-2 residual variance which is not related to the sample size at the individual level (*n*), while the MDESD formulas for a level-1 binary moderator with nonrandomly varying slope in Table 1 only includes the component associated with the level-1 residual variance. As a result, n is more influential on the MDESD than the MDES.

Figure 2 shows the relationship between power and cluster sample size by comparing the main treatment effect analysis with moderation analyses with binary level-1 and -2 moderators. The figure is based on the following assumptions: the intraclass correlation coefficient (ρ) is 0.2 in Figure 2a and 0.1 in Figure 2b in two-level CRTs. The proportions of variance explained by the covariates at level 1 and level 2 for the main effect analysis is 0.5 ($R_1^2 = R_2^2$ =0.5); The proportions of variance explained for the level-2 moderation analysis, R_2^2 = 0.5 at level 2, and for the level-1 moderation analysis, R_1^2 = 0.5 at the level 1. The proportion of variance between clusters on the effect of the student-level moderator explained by the school-level predictor is set to 0 (R_{2T}^2 = 0). The effect heterogeneity (ω) for the student-level moderator across school-levels is assumed as 0.3 for the randomly varying slope model, which is equivalent to an effect size variability of 0.06. We assume a balanced design with equal assignment of schools to the treatment and control groups (P = 0.5) and 20 students per school. In addition, half of the sample is in one moderator subgroup (Q = 0.5). For comparison purposes, we assume the effect size for the main treatment effect and the effect size difference for the moderator effect (standardized mean difference) to be detected using a two-sided test with α = 0.05 are both 0.20. This is equivalent to effect sizes for the two moderator subgroups of 0.3 and 0.1, respectively. The resulting power curves are for the moderation analyses with a binary level-2 moderator (grey solid line), a binary level-1 moderator with randomly varying slope (long dashed black line), a binary level-1 moderator with nonrandomly varying slope (short dotted black line), and the main treatment effect analysis (black solid line).

Figure 2 (a and b) indicates that the power increases for a binary level-1 moderator effect with the increase of the group sample size The power for detecting the effects of a binary level-1 moderator with nonrandomly varying slope (short dotted black line) is bigger than that for a binary level-1 moderator with randomly varying slope (long dashed black line). The power for detecting the effects of a binary level-1 moderator with nonrandomly varying slope (short dotted black line) is bigger than the power for the main treatment effect analysis (black solid line) in Figure 2a (ρ = 0.20). By comparing Figure 2a (ρ = 0.20) with Figure 2b (ρ = 0.10), we can see that the power for detecting the effect of a binary level-1 moderator with nonrandomly varying slope (short dotted black line) is bigger when the intraclass correlation is bigger. This is also apparent in the formulas for the MDESD which contain a factor of (1 - ρ), hence when ρ increases the MDESD decreases and the power increases. Note that across all scenarios the power for a binary level-2 moderator effect (grey solid line) is the smallest.

[Figure 2 about here]

Conclusion

The main findings are summarized as follows. First, the effects of the sample sizes at different levels, the levels of the moderators at which they have been assessed, the slopes of level-1 moderators (random vs. non-randomly varying), the distribution of moderators (binary vs. continuous), and the inclusion of covariates on power and MDESD in two-level CRTs are consistent with that in three-level CRTs (Dong, Spybrook, & Kelcey, 2018). For instance, the sample size at the higher level (e.g., level 2) is more critical than the sample size at lower level (e.g., level 1) for increasing the power to detect the effects of a level-2 moderator and a level-1 moderator with randomly varying slope. However, the sample size at level 1 is as important as at level 2 for increasing the power to detect the effect of a level-1 moderator with nonrandomly varying slope. Furthermore, the MDESD is larger or the power is smaller when the moderator is at the higher level. In other words, studies are more likely to be well-powered to detect level-1 moderator effects than level-2 moderator effects. Besides, the MDESD measured by the standardized mean difference for the binary moderator is always $1/\sqrt{Q(1-Q)}$ times of the MDESD measured by the standardized coefficient for the continuous moderator when it is level-2 moderator or level-1 moderator with nonrandomly varying moderator slope. In addition, including level-1 covariates can improve power for both level-1 and level-2 moderator effects; including level-2 covariates may improve power only if the level-2 covariates are in the intercept model for the level-2 moderator or the level-2 covariates are in the slope model to explain the heterogeneity of the level-1 moderator.

Second, when the estimation models are correctly specified for the real data, the model with a varying moderator slope will yield less precise estimates than the model with a constant moderator slope. The differences on the power and MDESD between the two models decreases when the number of clusters (J) increases and the effect heterogeneity (ω) decreases.

Lastly, the mismatch between the study design and real data will result in either overestimating or underestimating the power. Specifically, if the randomly varying slope model is used to design the studies where $\omega = 0$, the power will be underestimated; if the nonrandomly varying slope model is used to design the studies where $\omega > 0$, the power will be overestimated. The bias in power estimates due to model mismatch decreases when the sample size for the clusters (J) increases and the effect heterogeneity (ω) decreases. However, it is generally preferable to use the randomly varying slope model to design the cross-level moderation studies unless there is strong theory or prior knowledge that the slope of the lower level moderator does not vary across clusters.

This study focused on two-level CRTs. There are many important directions for further work. First, extending the work to other designs is necessary. This includes multisite randomized trials (MRTs), which are also common designs used to evaluate the effectiveness of programs (Spybrook, Shi, & Kelcey, 2016), and longitudinal study designs. Second, a well conducted power analysis heavily relies on accurate empirical estimates of the design parameters. Hence more empirical studies of design parameters such as the ICC, effect heterogeneity of level-1 covariates, and meaningful moderator effect size differences are important as we move forward.

Electronic Supplementary Material

- Electronic Supplementary Material 1.pdf

This pdf file contains the derivations of power and MDESD formulas.

- Electronic Supplementary Material 2.pdf

This pdf file contains the procedures and results of Monte Carlo simulation (Tables 1-24).

References

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* SAGE Publication, Inc.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In Howard S. Bloom (editor), *Learning more from social experiments: Evolving analytic approaches*, 115-172, 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., Richburg-Hayes, L. & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

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.

Kreft, G. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in Hierarchical Linear Models. *Multivariate Behavioral Research, 30* (1), 1-21.

LaHuis, D., Jenkin, D. R., Hartman, M. J., Hakoyama, S., & Clark, P. (in press). The effects of misspecifying the random part of multilevel models. *Methodology*.

Mathieu, J .E., Aguinis, H., Culpepper, S. A., and Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97* (5), 951-966.

Moerbeek, M. (2006). Power and money in cluster-randomized trials: when is it worth measuring a covariate? *Statistics in Medicine, 25*(15), 2607-2617.

Moerbeek, M., & Maas, C.J.M. (2005). Optimal experimental designs for multilevel logistic models with two binary predictors. *Communications in Statistics – Theory and Methods, 34*(5), 1151-1167.

Moerbeek, M., van Breukelen, G.J.P., & Berger, M.P.F. (2001). Optimal experimental designs for multilevel models with covariates. *Communications in Statistics, Theory and Methods, 30* (12), 2683-2697.

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

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis. 29* (1): 5-29.

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., 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* (3), 255-267. doi: 10.1080/1743727X.2016.1150454

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.* doi: 10.3102/1076998616655442

TABLE 1: Summary of standardized noncentrality parameters, MDESD and 100*(1-*α*)% confidence intervals for two-level CRTs

Model Number	HLM	Standardized Noncentrality Parameter (*λ*)	MDESD and 100*(1-*α*)% Confidence Interval	Degree of Freedom (*v*)
CRT2-1N	L1: $Y_{ij} = \beta_{0j} + \beta_{1j}S_{ij} + r_{ij}$, $r_{ij} \sim N(0, \sigma^2_{\mid S})$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + u_{0j}$, $u_{0j} \sim N(0, \tau^2_{\mid T})$ $\beta_{1j} = \gamma_{10} + \gamma_{11}T_j$	Binary Moderator: $\sqrt{\frac{\delta^2_{1b}P(1-P)Q(1-Q)Jn}{(1-R^2_1)(1-\rho)}}$ Continuous Moderator: $\sqrt{\frac{\delta^2_{1c}P(1-P)Jn}{(1-R^2_1)(1-\rho)}}$	Binary Moderator: $M_v\sqrt{\frac{(1-R^2_1)(1-\rho)}{P(1-P)Q(1-Q)Jn}}$ $(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R^2_1)(1-\rho)}{P(1-P)Q(1-Q)Jn}}$ Continuous Moderator: $M_v\sqrt{\frac{(1-R^2_1)(1-\rho)}{P(1-P)Jn}}$ $(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R^2_1)(1-\rho)}{P(1-P)Jn}}$	*J*(*n* -1)-2
CRT2-1R	L1: $Y_{ij} = \beta_{0j} + \beta_{1j}S_{ij} + r_{ij}$, $r_{ij} \sim N(0, \sigma^2_{\mid S})$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + u_{0j}$, $\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + u_{1j}$ $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{00\mid T} & \tau_{01\mid T} \\ \tau_{10\mid T} & \tau^2_{11\mid T} \end{pmatrix}\right)$	Binary Moderator: $\sqrt{\frac{\delta^2_{1b}P(1-P)J}{(1-R^2_{2T})\rho\omega + (1-R^2_1)(1-\rho)/(nQ(1-Q))}}$ Continuous Moderator: $\sqrt{\frac{\delta^2_{1c}P(1-P)J}{(1-R^2_{2T})\rho\omega + (1-R^2_1)(1-\rho)/n}}$	Binary Moderator: $M_v\sqrt{\frac{(1-R^2_{2T})\rho\omega + (1-R^2_1)(1-\rho)/(nQ(1-Q))}{P(1-P)J}}$ $(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R^2_{2T})\rho\omega + (1-R^2_1)(1-\rho)/(nQ(1-Q))}{P(1-P)J}}$ Continuous Moderator: $M_v\sqrt{\frac{(1-R^2_{2T})\rho\omega + (1-R^2_1)(1-\rho)/n}{P(1-P)J}}$ $(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R^2_{2T})\rho\omega + (1-R^2_1)(1-\rho)/n}{P(1-P)J}}$	*J* - 2
CRT2-2	L1: $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij}, r_{ij} \sim N(0, \sigma^2_{\mid X})$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01}S_j + \gamma_{02}T_j + \gamma_{03}(S_j \times T_j) + \gamma_{04}W_j + \gamma_{05}\bar{X}_{.j} + u_{0j}$, $u_{0j} \sim N(0, \tau^2_{\mid S,W,\bar{X},T})$ $\beta_{1j} = \gamma_{10}$	Binary Moderator: $\sqrt{\frac{\delta^2_{2b}P(1-P)Q(1-Q)(J-6)}{(1-R^2_2)\rho + (1-R^2_1)(1-\rho)/n}}$ Continuous Moderator: $\sqrt{\frac{\delta^2_{2c}P(1-P)(J-6)}{(1-R^2_2)\rho + (1-R^2_1)(1-\rho)/n}}$	Binary Moderator: $M_v\sqrt{\frac{(1-R^2_2)\rho + (1-R^2_1)(1-\rho)/n}{P(1-P)Q(1-Q)(J-6)}}$ $(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R^2_2)\rho + (1-R^2_1)(1-\rho)/n}{P(1-P)Q(1-Q)(J-6)}}$ Continuous Moderator: $M_v\sqrt{\frac{(1-R^2_2)\rho + (1-R^2_1)(1-\rho)/n}{P(1-P)(J-6)}}$ $(M_v \pm t_{\alpha/2})\sqrt{\frac{(1-R^2_2)\rho + (1-R^2_1)(1-\rho)/n}{P(1-P)(J-6)}}$	*J* – 6

Note. CRT2-1N and CRT2-1R stand for two-level CRTs with a level-1 moderator with nonrandomly varying and randomly varying slopes, respectively. CRT2-2 stands for two-level CRTs with a level-2 moderator

TABLE 2

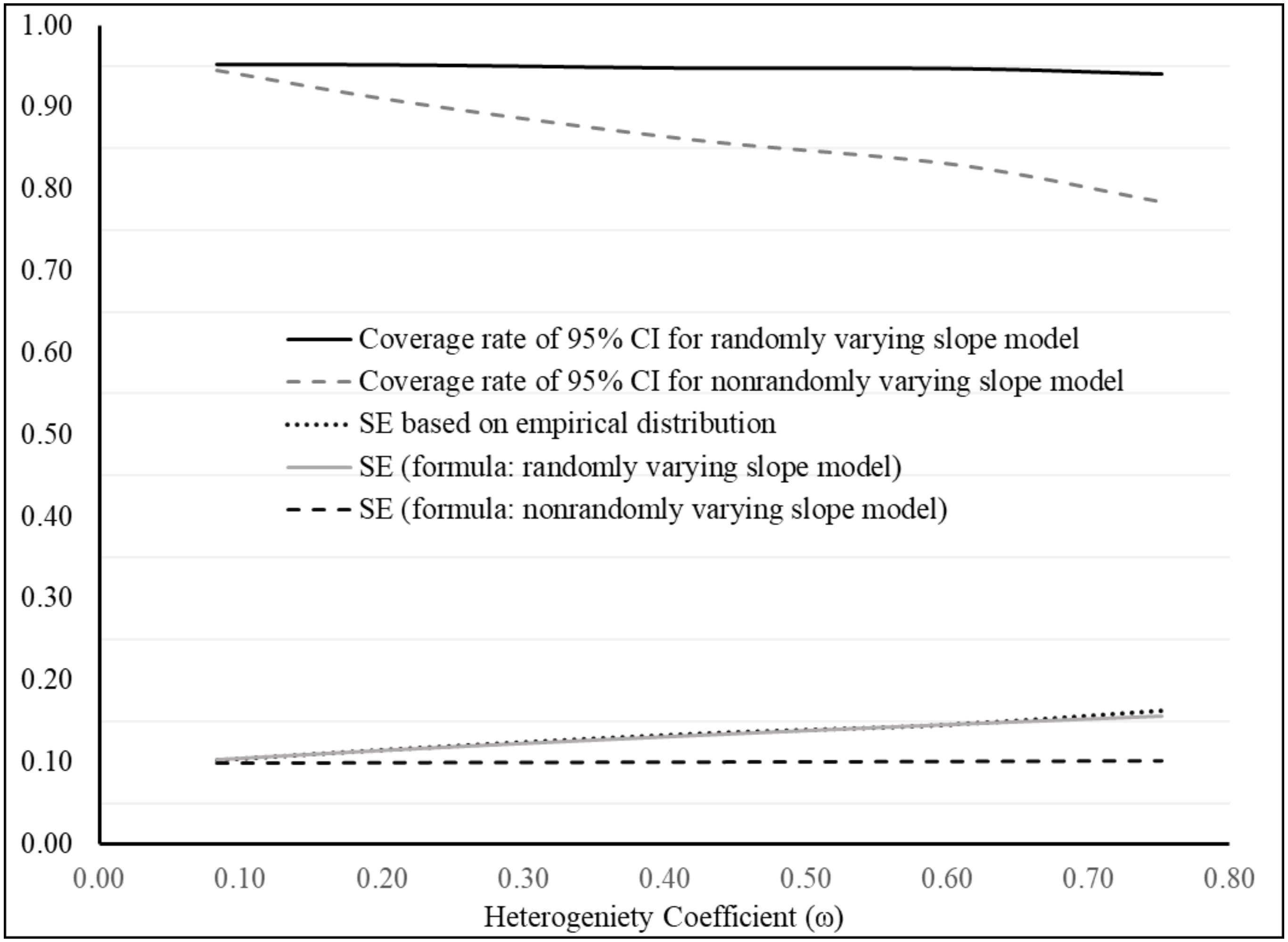
MDESD and statistical power of two-level CRTs

Level of Moderator	Slope of Lower Level Moderator	MDESD				Power			
		Binary Moderator		Continuous Moderator		Binary Moderator		Continuous Moderator	
		$J = 40$	$J = 80$	$J = 40$	$J = 80$	$J = 40$	$J = 80$	$J = 40$	$J = 80$
1	Nonrandomly Varying	0.11	0.08	0.06	0.04	1.00	1.00	1.00	1.00
1	Randomly Varying	0.26	0.18	0.25	0.17	0.56	0.86	0.63	0.91
2	NA	0.67	0.45	0.34	0.23	0.13	0.24	0.39	0.70

Note. Under the assumptions: $n = 100$, $\rho = 0.23$, $P = 0.5$, $Q = 0.5$, $R_1^2 = 0.5$, $R_2^2 = 0.5$, $R_{2T}^2 = 0$ and $\omega = 0.3$ for random slope design, power = 0.8 for the calculation of MDESD, and effect size difference = 0.2 for the calculation of power, a two-sided test with $\alpha = 0.05$.

Figure 1

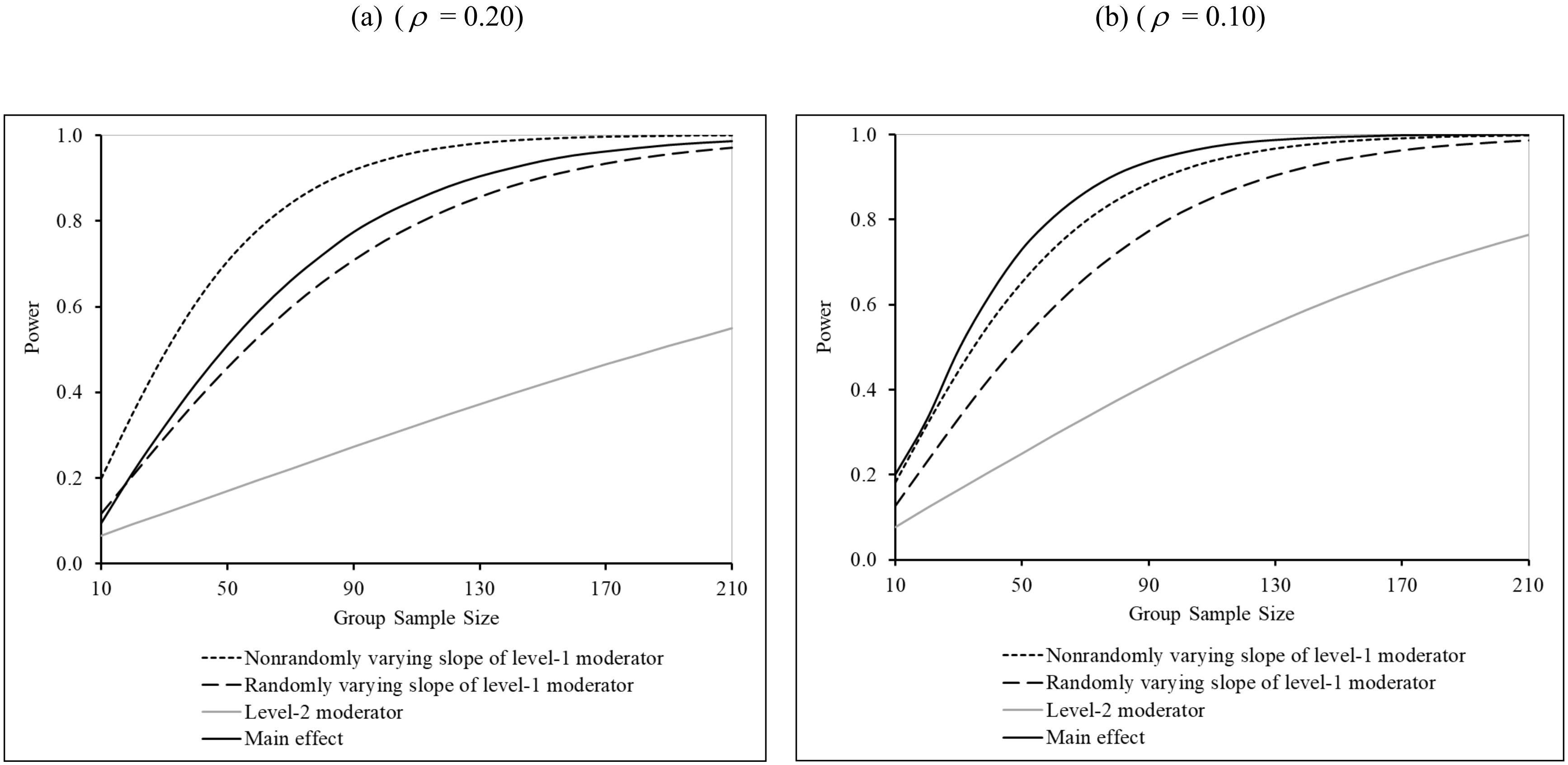
Standard Error (SE) and Coverage Rate of 95% CI vs. Heterogeneity Coefficient



Note. Under the assumptions: $\rho = 0.2$, $J = 40$, $n = 20$, $R_1^2 = 0.4$, $R_{2T}^2 = 0.07$, $P = 0.5$, $Q = 0.5$, effect size difference = 0.2.

Figure 2

Power vs. group sample size


(a) (ρ = 0.20)

(b) (ρ = 0.10)

Note. Under the assumptions: $n = 20$, $R_1^2 = 0.5$, $R_2^2 = 0.5$, $P = 0.5$, $Q = 0.5$, $R_{2T}^2 = 0$ and $\omega = 0.3$ for randomly varying slope design, effect size (standardized mean difference) = 0.2, effect size difference (standardized mean difference) = 0.2, and a two-sided test with $\alpha = 0.05$.