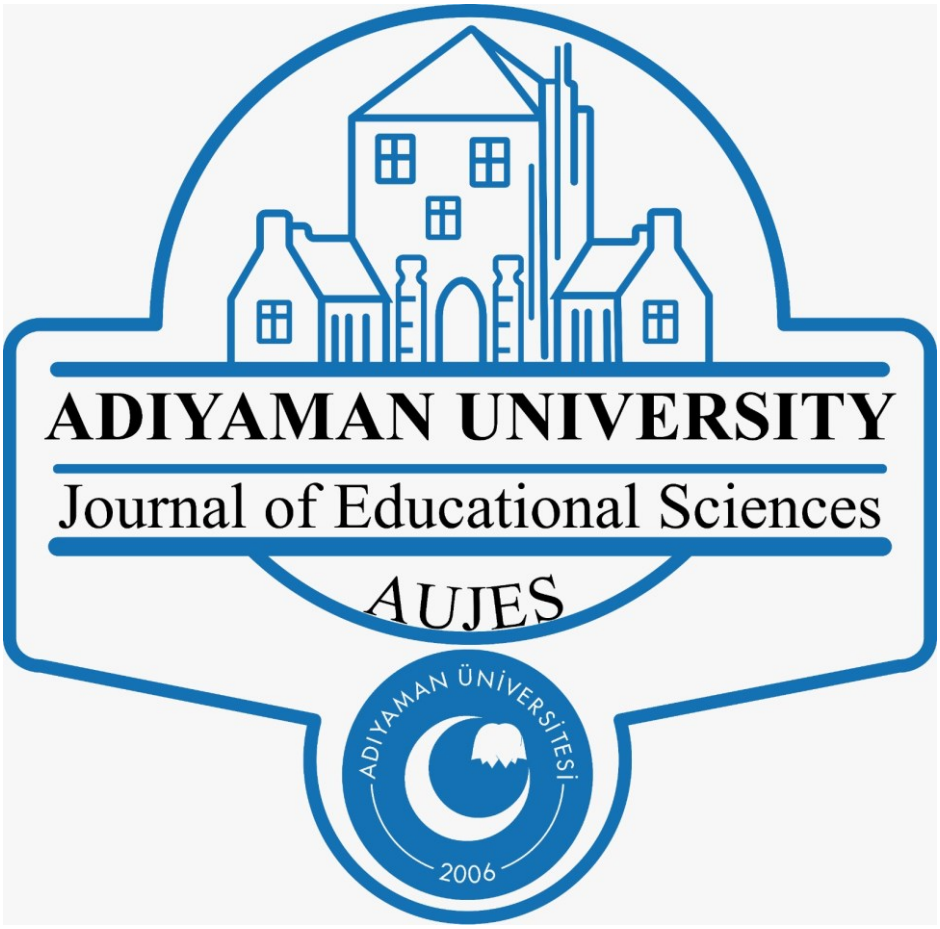

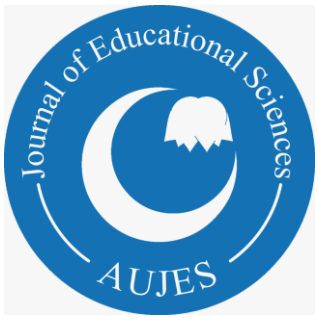

ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

https://dergipark.org.tr/tr/pub/adyuebd

Consequences of Ignoring a Level of Nesting on Design and Analysis of Blocked Three-level Regression Discontinuity Designs: Power and Type I Error Rates

Metin Bulus
Adiyaman University
Nianbo Dong
University of North Carolina – Chapel Hill

To cite this article:

Bulus, M., & Dong, N. (2022). Consequences of ignoring a level of nesting on design and analysis of blocked three-level regression discontinuity designs: Power and Type I error rates. *Adiyaman University Journal of Educational Sciences*, *Vol*(No), Page X-Page Y.

Consequences of Ignoring a Level of Nesting on Design and Analysis of Blocked Three-level Regression Discontinuity Designs: Power and Type I Error Rates*

Metin Bulus**
Adiyaman University

Nianbo Dong
University of North Carolina – Chapel Hill

Abstract

Multilevel regression discontinuity designs have been increasingly used in education research to evaluate the effectiveness of policy and programs. It is common to ignore a level of nesting in a three-level data structure (students – classrooms/teachers – schools), whether due to resource constraints during the planning phase or unwittingly during data analysis. This study aims to investigate consequences of ignoring either intermediate or top-level in blocked three-level regression discontinuity (BIRD3) designs during data analysis and planning. During analysis, results indicated that ignoring a level did not affect treatment effect estimates; however, it affected power and Type I error rates. Ignoring intermediate level did not cause a significant problem. Power rates were slightly underestimated, whereas Type I error rates were stable. In contrast, ignoring a top-level resulted in high power rates, however, severe inflation in Type I error deemed this strategy ineffective. During planning, it is viable to use parameters from a misspecified two-level blocked regression discontinuity model where level 2 was ignored (BIRD2_L2_Ignored) for a future BIRD2 design. They can also be used for a future BIRD3 design where level 2 (top level) design parameters in the BIRD2_L2_Ignored model are substituted for level 3 design parameters. However, when level 2 (top level) design parameters in the BIRD2_L2_Ignored model are used for level 2 (intermediate level) design parameters in a future BIRD3 design, researchers risk having low power rates after data collection and analysis. Ignoring the top-level (BIRD2_L3_Ignored) was more problematic. Calculated power rates were unstable; thus, using parameters from BIRD2_L3_Ignored model in a future BIRD2 or BIRD3 designs should be avoided.

Keywords: blocked regression discontinuity designs, ignoring a level of nesting, power analysis, sample size, minimum detectable effect size

Introduction

One of the fundamental assumption of Ordinary Least Squares (OLS) regression is the independence of observations. This assumption is violated when errors are not independent of each other (presenting autocorrelation) due to nesting of observations within organizational structures (Bickel, 2007; Finch & Bolin, 2017; Goldstein, 2011; Hox, 2010; Raudenbush & Bryk, 2002; Snijder & Bosker, 2011). Violation of independence presents challenges to hypothesis testing. It is well known that bias in point estimate is ignorable, but OLS regression produces overly optimistic standard errors, leading to inflated Type I errors (Finch & Bolin, 2017; Singer, 1987; Fox, 1997). Multilevel linear modeling (MLM) arouse as a compelling option for remedying the violation of independent errors in the case where nesting structure consists of mutually exclusive groups (such as classrooms, teachers, or schools in education systems). Additionally, MLM allows inspection of more complex research questions. One can study the influence of contextual factors on the outcome of interest and the estimates of predictors. The latter can be translated into substantial research questions on treatment effect heterogeneity and cross-level interactions. In the past 30 years, MLM has been prevalently used in education research to answer substantial research questions owing to rapid advances in its

* This article was produced from corresponding author's doctoral dissertation titled "Design Considerations in Three-level Regression Discontinuity Studies" from University of Missouri – Columbia. This project has been funded by the National Science Foundation (DGE-1913563). The opinions expressed herein are those of the authors and not the funding agency.
** Corresponding Author, Metin Bulus, bulusmetin@gmail.com, ORCID: 0000-0003-4348-6322

methodology, development of publicly available software, and accessible literature (e.g., Bickel, 2007; Finch & Bolin, 2017; Goldstein, 2011; Hox, 2010; Raudenbush & Bryk, 2002; Snijder & Bosker, 2011, among many others).

However, the complex structure of the education system presents challenges to data collection efforts. Data collection efforts on all levels of organizations and actors (students, teachers, administrators, schools, and states) are partially hindered by lack of economic resources, lack of administrative records, or partially by researchers via unwittingly ignoring what could matter. In one scenario, a researcher could collect data from only students, in the other, from students and classrooms/teachers but not schools, yet in another, from students and schools but not classrooms/teachers. In other words, one of the levels in the organizational structure (e.g., classroom/teachers or schools) could be ignored or omitted. The omission of intermediate level (classrooms/teachers) is typical in practice, sometimes due to the absence of administrative records that identify which classroom or teacher the child belongs (Zhu et al., 2011), or due to simplicity or small sample sizes (van Den Noorthgate et al., 2005). In education, the most common version of ignoring a nesting level occurs when classroom level information is ignored. However, the proportion of variance attributed to classroom level can exceed that of school level (Goldstein, 2011; Muthen, 1991), or the magnitude of this variance can be subject-specific. For instance, the proportion of variance in the mathematic achievement attributed to classroom level is higher than the proportion of variance in the reading achievement compared to the school level variance (Nye et al., 2004; Raudenbush & Bryk, 2002). Despite the possibility of a sizeable proportion of variance attributed to the intermediate level, many empirical studies did not acknowledge classroom level information in the analysis (e.g., Konu et al., 2002; Raudenbush & Bryk, 1986). Some recent evaluation studies indicate that regression discontinuity designs (RDDs) are not exempt from this problem (see Jenkins et al., 2016; Konstantopoulos & Shen, 2016, Luyten, 2006; May et al., 2016). The literature consistently demonstrated that ignoring a top or intermediate level has a detrimental effect on variance components, estimates, and standard errors. Some studies reported the effect of ignoring a nesting level on variance components (Moerbeek, 2004; Opdenakker & van Damme, 2000), while some studies focused on both variance components and standard errors (van Den Noortgate et al., 2005; Zhu et al., 2011). From this point forward, for brevity, we will refer to level 1 as L1, level 2 as L2 and level 3 as L3.

Effects of Ignoring a Level of Nesting on Variance Components

Using a three-level model (students as L1– classrooms/teachers as L2 – schools as L3), in the case of a balanced design[†], Moerbeek (2004) found that ignoring L3 did not affect the variance component at L1 but inflated the variance component at L2. The amount of inflation in the variance at L2 was approximately equal to the ignored amount at L1. Similarly, using a four-level model (students as L1 – teachers as L2 – classrooms as L3 – schools as L4), van Den Noortgate et al. (2005) concluded that omission of L4 did not affect variance estimates at L2 and L1. However, the ignored variance at L4 was transferred to the variance at L3.

The consequences of ignoring an intermediate level are more complicated than ignoring the top level. van Den Noortgate et al. (2005) found that the omission of an intermediate level (L2 or L3 in a four-level model) resulted in inflation of the variance estimates at the flanking levels. For example, if L3 was omitted, the variance was distributed to L2 and L4, which confirms findings by Moerbeek (2004) and Opdenakker and van Damme (2000). Moerbeek (2004) noted that inflation in variance components depended on the magnitude of the variance component at the ignored level, the level at which predictor variable was measured, and sample sizes at one or more levels.

Effects of Ignoring a Level of Nesting on Standard Errors

The literature already established that fixed effect estimates themselves are not affected as much when one relies on OLS estimation instead of MLM, whereas standard errors are overly optimistic (Finch & Bolin, 2017; Singer, 1987; Fox, 1997). If one relies on OLS estimation instead of MLM in the face of a multilevel data structure, it implies that all levels of nesting are ignored. When variance component of a given level is affected due to ignoring of a level of nesting, naturally, standard errors of the estimates at that level and those at the ignored level could also be affected (Opdenakker & Van Damme, 2000).

In the case of a balanced design, using a three-level model (students as L1– classrooms as L2– schools as L3), Moerbek (2004) found that inflation in standard errors depended on the ignored level (L2 versus L3), the level at which predictor variable was measured, the magnitude of the proportion of variance attributed to ignored level, and sample sizes at one more level. For example, ignoring L2 inflates standard errors for the fixed effect estimates at L1, resulting in inflated *p*-values but not those at L3 (Moerbek, 2004). However, as Moerbek (2004) noted, if the proportion of variance

[†] This article was produced from corresponding author's doctoral dissertation titled "Design Considerations in Three-level Regression Discontinuity Studies" from University of Missouri

attributed to the ignored level was minor, standard errors of fixed effect estimates were not affected to a great extent. This finding was later confirmed by Zhu et al. (2011) using elementary school data.

Using a four-level model (students as L1– teachers as L2 - classrooms as L3– schools as level 4), van Den Noortgate et al. (2005) found that, in general, the standard error of the intercept and estimates at the ignored or adjacent levels were affected. When level 4 was ignored, the standard error of the estimate for predictors at L3 was affected. When L3 was ignored in a balanced data, the standard error of the estimate for predictors at L2 increased. In contrast, the standard error of the intercept and estimates for predictors at the ignored level decreased. When the data was unbalanced, however, the standard error of the estimates for predictors at level 4 decreased when L3 was ignored.

Opdenakker and Van Damme (2000) found that regardless of the level ignored, the standard error of the intercept was underestimated. However, when level 4 was ignored, the standard error of the estimates at levels 1 and 2 was not affected as much. If the predictor itself belongs to the ignored level, then the standard error of their estimates was underestimated. Zhu et al. (2011) extended previous work on ignoring a nesting structure by mainly focusing on the design phase of cluster-randomized trials rather than analysis, although results apply to both. In particular, authors considered design parameters from two-level data to design three-level studies. Manipulating and analyzing four empirical multi-site datasets (including elementary and secondary school data), Zhu et al. (2011) concluded that ignoring the intermediate level had no substantial effects on statistical power, precision or standard error of the estimate for predictors at L3. Additionally, they concluded that using design parameters from a two-level study to design a three-level study did not pose a substantial threat.

Evidence from Empirical Studies that Ignore a Level of Nesting in RDD

Several studies from 2000 onward focused on the cutoff-based assignment at the individual level, which, one way or another, were adjusting estimates for clustering (or nesting structure). About a quarter of these studies adjusted for clustering effects using MLM framework (Hustedt et al., 2015; Luyten, 2006; Luyten et al., 2008; May et al., 2016), and about a quarter of the studies used Lee and Card (2008) method (Balu et al., 2015; Cortes, 2015; Deke et al., 2012; Harrington et al., 2016; Reardon et al., 2010). The remaining studies either used bootstrap methods or none (Jenkins et al., 2016; Klerman et al., 2015; Leeds et al., 2017; Ludwig & Miller, 2005; Matsudarie, 2008; Wong et al., 2008). The four RDDs relying on individual level cutoff-based assignment and the MLM framework are summarized below.

Hustedt et al. (2015) evaluated the effectiveness of the Arkansas Better Chance (ABC) initiative at kindergarten on student achievement, relying on the state's strict age-based admission criteria to the program. Although they analyzed the data using single-level analysis, district-level information was included in the model as fixed effects. Luyten et al. (2008) used Progress in International Reading Literacy Study (PIRLS) 2000 large-scale assessment data to examine the effect of an extra year of schooling on student achievement relying on the cutoff that split students into 9^{th} and 10^{th} grades. They analyzed the data using a two-level model where the schooling effect was assumed to vary across schools. Luyten (2006) used Trends in International Mathematics and Science Study (TIMSS) 1995 large-scale assessment data to examine the effect of an extra year of schooling on student achievement, relying on the cutoff that split students into consecutive grades. Similar to Luyten et al. (2008), a two-level model was used where the schooling effect is assumed to vary. May et al. (2016) evaluated the effectiveness of Reading Recovery i3 Scale-Up on students' achievement in first and third grades relying on students' pretest scores. They analyzed the data using a two-level RDD where the program effect was assumed to vary across schools. In summary, four RDD studies relying on individual level cutoff-based assignment and also used MLM framework could have been analyzed by acknowledging the classroom level information or district or state-level fixed effects.

Problem Statement

Drawing from four multi-site empirical elementary and secondary school datasets, Zhu et al. (2011) concluded that using design parameters from a two-level study for a future three-level design did not create a substantial problem. However, scholars in school effectiveness research portray a different picture (Moerbek, 2004; Opdenakker & van Damme, 2000; van Der Noortgate et al., 2005). Unlike Zhu et al. (2011), these scholars usually focused on the data analysis phase. The effect of using design parameters from a two-level study for a three-level design is less known when the treatment variable is at L1. In this study, within the context of blocked two-level RDD (BIRD2) and blocked three-level RDD (BIRD3), we investigate whether it is plausible to use design parameters from a misspecified BIRD2 model (where intermediate or top-level in BIRD3 design is ignored) for a future BIRD3 design. Specifically, we investigate the following questions:

1. How do variance components shift when intermediate or top-level in a BIRD3 model is ignored?
2. How standard error of the treatment effect estimate (a L1 predictor) is affected by these misspecifications?

3. Can we use design parameters from a misspecified BIRD2 model (where intermediate or top-level in BIRD3 design is ignored) for a future BIRD2 or BIRD3 design?

Method

Consider a nested sampling structure consisting of three levels (e.g., students at L1 – classrooms at L2 – schools at L3), with an assignment variable *S* and a predetermined cutoff S_0 at L1 (from which treatment variable *T* is derived), a covariate *X* at L1, a covariate *W* at L2, and a covariate *V* at L3. Assume intercepts and treatment effect is random across L2 and 3 units. Also, assume that the data is balanced, that is, *n* number of L1 units per L2 unit, *J* number of L2 units per L3 unit, and *K* number of L3 units. Balanced data is not the requirement for the model or the estimation procedure; however, statistical power of the average treatment effect estimate (obtained from the data) approximates formula-based power calculations in the `cosa` R package (Bulus & Dong, 2021a; Bulus & Dong, 2021b) and PowerUp! software (Dong & Maynard, 2013).

Statistical Models

The following models pertain to the analysis of correctly specified BIRD3 model.

Unconditional Model

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

L1: $Y_{ij} = \beta_{0jk} + r_{ijk}$

L2: $\beta_{0jk} = \gamma_{00k} + \mu_{0jk}$

L3: $\gamma_{00k} = \xi_{000} + \varsigma_{00k}$,

where $r_{ijk} \sim N(0, \sigma^2)$, $\mu_{0jk} \sim N(0, \tau_2^2)$ and $\varsigma_{00k} \sim N(0, \tau_3^2)$.

Treatment Only Model

The following model is used to obtain variance parameters τ_{T2}^2 and τ_{T3}^2, as defined below, which will be used to calculate various standardized parameters along with parameters from unconditional and full models.

L1: $Y_{ij} = \beta_{0jk} + \beta_{1jk} T_{ijk} + r_{ijk}$

L2: $\beta_{0jk} = \gamma_{00k} + \mu_{0jk}$

$\beta_{1jk} = \gamma_{10k} + \mu_{1jk}$

L3: $\gamma_{00k} = \xi_{000} + \varsigma_{00k}$

$\gamma_{10k} = \xi_{100} + \varsigma_{10k}$,

where $r_{ijk} \sim N(0, \sigma_{|T}^2)$, $\begin{pmatrix}\mu_{0jk}\\ \mu_{1jk}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_{2|T}^2 & \tau_{2T2}\\ \tau_{2T2} & \tau_{T2}^2\end{pmatrix}\right)$ and $\begin{pmatrix}\varsigma_{00k}\\ \varsigma_{10k}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_{3|T}^2 & \tau_{3T3}\\ \tau_{3T3} & \tau_{T3}^2\end{pmatrix}\right)$.

Full Model

The following model is used to generate data for Monte Carlo simulations. It is also used to obtain variance parameters $\sigma_{|X}^2$, $\tau_{2|W}^2$, and $\tau_{3|V}^2$, as defined below, which are used to calculate various standardized parameters along with the parameters from unconditional and treatment only model. In addition to estimation of the treatment effect, empirical standard error and empirical power rates are estimated using this model.

L1: $Y_{ij} = \beta_{0jk} + \beta_{1jk} T_{ijk} + \beta_{2jk}(S_{ijk} - S_0) + \beta_{3jk} X_{ijk} + r_{ijk}$

L2: $\beta_{0jk} = \gamma_{00k} + \gamma_{01k} W_{jk} + \mu_{0jk}$

$\beta_{1jk} = \gamma_{10k} + \gamma_{11k} W_{jk} + \mu_{1jk}$

$\beta_{2jk} = \gamma_{20k}$

$$\beta_{3jk} = \gamma_{30k}$$

L3: $\gamma_{00k} = \xi_{000} + \xi_{001}V_k + \varsigma_{00k}$

$$\gamma_{10k} = \xi_{100} + \xi_{101}V_k + \varsigma_{10k}$$

$$\gamma_{20k} = \xi_{200}$$

$$\gamma_{30k} = \xi_{300}$$

$$\gamma_{01k} = \xi_{010}$$

$$\gamma_{11k} = \xi_{110},$$

where $r_{ijk} \sim N(0, \sigma^2_{|X})$, $\begin{pmatrix}\mu_{0jk}\\ \mu_{1jk}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau^2_{2|W} & \tau_{2T2|W}\\ \tau_{2T2|W} & \tau^2_{T2|W}\end{pmatrix}\right)$ and $\begin{pmatrix}\varsigma_{00k}\\ \varsigma_{10k}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau^2_{3|V} & \tau_{3T3|V}\\ \tau_{3T3|V} & \tau^2_{T3|V}\end{pmatrix}\right)$ and where

$\rho_2 = \frac{\tau_2^2}{\tau_3^2+\tau_2^2+\sigma^2}$ and represents proportion of variance in the outcome between L2 units,

$\rho_3 = \frac{\tau_3^2}{\tau_3^2+\tau_2^2+\sigma^2}$ and represents proportion of variance in the outcome between L3 units,

$\omega_2 = \frac{\tau_{T2}^2}{\tau_2^2}$ and represents treatment effect heterogeneity across L2 units,

$\omega_3 = \frac{\tau_{T3}^2}{\tau_3^2}$ and represents treatment effect heterogeneity across L3 units,

σ^2 is the L1 variance,

$\tau^2_{3|V}$ is the L3 variance conditional on L3 variables,

$\tau^2_{2|W}$ is the L2 variance conditional on L2 variables,

$R_1^2 = 1 - \sigma^2_{|X}/\sigma^2$ and is the L1 variance explained by L1 variables,

$R^2_{T2} = 1 - \tau^2_{T2|W}/\tau^2_{T2}$ and is the proportion of variance at L2 on the treatment explained by L2 variables,

$R^2_{T3} = 1 - \tau^2_{T3|V}/\tau^2_{T3}$ and is the proportion of variance at L3 on the treatment explained by L3 variables.

Standard Error for Correctly Specified BIRD3

For the correctly specified BIRD3 model, standard error of the treatment effect takes the form of (Bulus & Dong, 2022)

$$SE(\hat{\xi}_{100}) = \sqrt{\frac{\omega_3\rho_3(1-R^2_{T3})}{K} + \frac{\omega_2\rho_2(1-R^2_{T2})}{KJ} + \frac{(1-\rho_3-\rho_2)(1-R_1^2)(RDDE)}{KJnp(1-p)}}$$

where RDDE is regression discontinuity design effect and takes the form of $RDDE = 1/(1-\rho^2_{TS})$ when only linear form of the score variable is added to the model (Bulus, 2022; Bulus & Dong, 2022; Schochet, 2008, 2009). ρ^2_{TS} is the squared correlation between treatment and score variables and defined as $\rho^2_{TS} = \sigma_{TS}/(\sqrt{p(1-p)\sigma_S})$ where σ_{TS} is the covariance between *T* and *S*, and σ_S is the standard deviation of *S* (see Bulus, 2022; Bulus & Dong, 2022; Schochet, 2008, 2009).

Standard Errors for Misspecified BIRD2 Model

When intermediate level in BIRD3 model is ignored, standard error of the treatment effect for the new BIRD2 model takes the form of (Bulus & Dong, 2022; Schochet, 2008, 2009)

$$SE(\hat{\xi}_{100}) = \sqrt{\frac{\omega_2\rho_2(1-R^2_{T2})}{K} + \frac{(1-\rho_2)(1-R_1^2)(RDDE)}{KJnp(1-p)}}$$

Different from the correctly specified BIRD3 model, ω_2 is now the treatment effect heterogeneity across L2 units (schools) in the misspecified BIRD2 model where only treatment variable is included, ρ_2 is the proportion of variance

in the outcome that is between L2 units (schools) in the unconditional misspecified BIRD2 model. Sample size for the top level remains K, however, the sample size at L1 is now Jn.

When top level in BIRD3 model is ignored, standard error of the treatment effect for the new BIRD2 model takes the form of (Bulus & Dong, 2022; Schochet, 2008, 2009)

$$SE(\hat{\xi}_{100}) = \sqrt{\frac{\omega_2\rho_2(1-R_{T2}^2)}{KJ} + \frac{(1-\rho_2)(1-R_1^2)(RDDE)}{KJnp(1-p)}}$$

Different from correctly specified BIRD3 model, ω_2 is now the treatment effect heterogeneity across L2 units (classrooms/teachers) in the misspecified BIRD2 model where only treatment variable is included, ρ_2 is the proportion of variance in the outcome that is between L2 units (classrooms/teachers) in the unconditional misspecified BIRD2 model. Sample size for the top level is now KJ, whereas the sample size at L1 remains n.

Monte Carlo Simulation

Population Parameters and Scenarios

We generated $S, X, W, V \sim N(0,1)$ and derived T from S and S_0 such that $p = 0.5$ or 0.2. Coefficients were manipulated such that ρ_2 and ρ_3 values are close to those commonly encountered in education settings. The two scenarios that produce different values of ρ_2 and ρ_3 are as follows (approximately ~0.40 and ~0.20 for Scenario 1 and ~0.15 and ~0.10 for Scenario 2):

Scenario 1

L1: $Y_{ij} = \beta_{0jk} + \beta_{1jk}T_{ijk} + 0.5(S_{ijk} - S_0) + 0.5X_{ijk} + r_{ijk}$

L2: $\beta_{0jk} = \gamma_{00k} + 0.3W_{jk} + \mu_{0jk}$

$\beta_{1jk} = \gamma_{10k} + 0.3W_{jk} + \mu_{1jk}$

L3: $\gamma_{00k} = 0 + 0.25V_k + \varsigma_{00k}$

$\gamma_{10k} = \xi_{100} + 0.25V_k + \varsigma_{10k}$,

where $r_{ijk} \sim N(0,1)$, $\begin{pmatrix}\mu_{0jk}\\ \mu_{1jk}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.5 & 0\\ 0 & 1.5\end{pmatrix}\right)$ and $\begin{pmatrix}\varsigma_{00k}\\ \varsigma_{10k}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 0.5\end{pmatrix}\right)$.

Scenario 2

L1: $Y_{ij} = \beta_{0jk} + \beta_{1jk}T_{ijk} + 0.3(S_{ijk} - S_0) + 0.3X_{ijk} + r_{ijk}$

L2: $\beta_{0jk} = \gamma_{00k} + 0.25W_{jk} + \mu_{0jk}$

$\beta_{1jk} = \gamma_{10k} + 0.25W_{jk} + \mu_{1jk}$

L3: $\gamma_{00k} = 0 + 0.2V_k + \varsigma_{00k}$

$\gamma_{10k} = \xi_{100} + 0.2V_k + \varsigma_{10k}$,

where $r_{ijk} \sim N(0,3)$, $\begin{pmatrix}\mu_{0jk}\\ \mu_{1jk}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.5 & 0\\ 0 & 1\end{pmatrix}\right)$ and $\begin{pmatrix}\varsigma_{00k}\\ \varsigma_{10k}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 0.5\end{pmatrix}\right)$.

Along with the four scenarios (Scenario 1 or 2, by $p = 0.5$ or 0.2) above, we determined treatment effect as $\xi_{100} = 0.25$ for statistical power simulation and as $\xi_{100} = 0$ for Type I error simulation. Additionally, we differed sample size $K = 50$ or 100, and kept $n = 20$ & $J = 5$ constant across all the scenarios. Sample sizes were chosen to approximate those commonly encountered in education. Although $J = 5$ may not be as common, to obtain consistent variance estimates it is an ideal minimum number. In total, there were eight scenarios for statistical power simulation (P1-P8) and eight scenarios for Type I error simulation (T1-T8).

Analysis

The data were generated for these eight (P1-P8 and T1-T8) scenarios using parameters described in the equations (see *Monte Carlo Simulation* section). As for the correctly specified model, each generated data set was analyzed using "Null Model," "Treatment Only Model," and "Full Model." We used PROC MIXED in SAS with default restricted maximum likelihood (REML) estimation and unstructured (UN) variance-covariance structure. For each scenario, the procedure was replicated 5000 times. Monte Carlo-based standard error (SE_{MC}) was calculated as the standard deviation of the 5000 treatment effect estimates. Monte Carlo-based power and Type I error rates were calculated based on the proportion of replications rejecting the null with a p-value smaller than 0.05. Other estimated parameters were averaged over 5000 replications. The standardized parameters that were used for power calculations are based on the averages over 5000 replications. There were 5000 rows for estimates, standard errors, and variance parameters, but only their averages were used to obtain standardized parameters.

Power Calculations

Averages were transformed into standardized parameters according to definitions in "Null Model," "Treatment Only Model," and "Full Model" described in the earlier section. Then, the standardized parameters were used in `power.bird3()` and `power.bird2()` functions for power calculation using `cosa` R library (Bulus & Dong, 2021a, 2021b). Model parameters, corresponding arguments, and their possible range are defined in Tables 1 and 2.

Table 1

BIRD3 Model Parameters,, Corresponding `cosa` *R Package Arguments and Their Range*

Parameter	$ES = \frac{\xi_{100}}{\sqrt{\tau_3^2 + \tau_2^2 + \sigma^2}}$	$\rho_2 = \frac{\tau_2^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$	$\rho_3 = \frac{\tau_3^2}{\tau_3^2 + \tau_2^2 + \sigma^2}$	$\omega_2 = \frac{\tau_{T2}^2}{\tau_2^2}$	$\omega_3 = \frac{\tau_{T3}^2}{\tau_3^2}$
`power.bird3()`	`es`	`rho2`	`rho3`	`omega2`	`Omega3`
Range	$ES \sim N(0,1)$	[0,1]	[0,1]	[0,1]	[0,1]
Parameter	g_3: number of L3 covariates excluding treatment	$R_1^2 = 1 - \frac{\sigma_{\|X}^2}{\sigma^2}$	$R_{T2}^2 = 1 - \frac{\tau_{T2\|W}^2}{\tau_{T2}^2}$	$R_{T3}^2 = 1 - \frac{\tau_{T3\|V}^2}{\tau_{T3}^2}$	p: proportion of subjects below (or above) the cutoff
`power.bird3()`	`g3`	`r21`	`r2t2`	`r2t3`	`p`
Range	$g_3 \in N^+$	[0,1]	[0,1]	[0,1]	(0,1)
Parameter	n_1	n_2	n_3		
`power.bird3()`	`n1`	`n2`	`n2`		
Range	$n_1 \in N^+$	$n_2 \in N^+$	$n_3 \in N^+$		

Table 2

BIRD2 Model Parameters, Corresponding `cosa` *R Package Arguments and Their Range*

Parameter	$ES = \frac{\gamma_{10}}{\sqrt{\tau_2^2 + \sigma^2}}$	$\rho_2 = \frac{\tau_2^2}{\tau_2^2 + \sigma^2}$	$\omega_2 = \frac{\tau_{T2}^2}{\tau_2^2}$	g_2: number of L2 covariates excluding treatment	$R_1^2 = 1 - \frac{\sigma_{\|X}^2}{\sigma^2}$
`power.bird2()`	`es`	`rho2`	`omega2`	`g2`	`r21`
Range	$ES \sim N(0,1)$	[0,1]	[0,1]	$g_2 \in N^+$	[0,1]
Parameter	$R_{T2}^2 = 1 - \frac{\tau_{T2\|W}^2}{\tau_{T2}^2}$	p: proportion of subjects below (or above) the cutoff	n_1	n_2	
`power.bird2()`	`r2t2`	`p`	`n1`	`n2`	

Range	[0,1]	(0,1)	$n_1 \in N^+$	$n_2 \in N^+$

For planning a BIRD2 design (planned-BIRD2) based on parameters from a misspecified BIRD2 where L2 was ignored (analyzed-msL2ig-BIRD2), we used `power.bird2()` function. Note that sample size at L1 for BIRD2 is the product of sample size at L1 and 2 in BIRD3. An example code chunk is presented below.

```
# misspecified BIRD2 (L2 ignored)
power.bird2(es = 0.10, rho2 = .30, omega2 = .54,
            g2 = 1, r21 = 0.22, r2t2 = 0.08,
            p = 0.50, n1 = 100, n2 = 50)
```

For planning a BIRD3 design (planned-BIRD3) based on parameters from a misspecified BIRD2 where L2 was ignored (analyzed-msL2ig-BIRD2), we used `power.bird3()` function. We assume that a researcher substituted L2 parameters in analyzed-msL2ig-BIRD2 for L3 parameters in planned-BIRD3. Thus, `rho2 = 0`, `omega2 = 0`, and `r2t2 = 0`. An example code chunk is presented below.

```
# L2 parms for analyzed-BIRD2 (L2 of the original BIRD3 ignored)
# can be substituted for L3 parms in planned-BIRD3
power.bird3(es = 0.10, rho2 = 0, rho3 = .30, omega2 = 0, omega3 = .54,
            g3 = 1, r21 = 0.22, r2t2 = 0, r2t3 = 0.08,
            p = 0.50, n1 = 20, n2 = 5, n3 = 50)
```

We can also assume that a researcher substituted L2 parameters in analyzed-msL2ig-BIRD2 for L2 parameters in planned-BIRD3. Thus, `rho3 = 0`, `omega3 = 0`, and `r2t3 = 0` as in the following.

```
# L2 parms for analyzed-BIRD2 (L2 of the original BIRD3 ignored)
# can be substituted for L2 parms in planned-BIRD3
power.bird3(es = 0.10, rho3 = 0, rho2 = .30, omega3 = 0, omega2 = .54,
            g3 = 0, r21 = 0.22, r2t3 = 0, r2t2 = 0.08,
            p = 0.50, n1 = 20, n2 = 5, n3 = 50)
```

Another possible scenario is that a researcher may attempt planning a BIRD2 design (planned-BIRD2) based on parameters from a misspecified BIRD2 where L3 was ignored (analyzed-msL3ig-BIRD2). Again, we used `power.bird2()` function. Note that, different from the previous scenario, the sample size at L2 for BIRD2 is the product of sample size at levels 2 and 3 in BIRD3. An example code chunk is presented below.

```
# misspecified BIRD2 (L3 ignored)
power.bird2(es = 0.10, rho2 = .61, omega2 = .65,
            g2 = 1, r21 = 0.53, r2t2 = 0.04,
            p = 0.50, n1 = 20, n2 = 250)
```

For planning a BIRD3 design (planned-BIRD3) based on parameters from a misspecified BIRD2 where L3 was ignored (analyzed-msL3ig-BIRD2), we used `power.bird3()` function. We assume that a researcher substituted L2 parameters in analyzed-msL2ig-BIRD2 for L3 parameters in planned-BIRD3. Thus, `rho2 = 0`, `omega2 = 0`, and `r2t2 = 0`. An example code chunk is presented below.

```
# L2 parms for the analyzed-BIRD2 (L3 of the original BIRD3 ignored)
# can be substituted for L3 parms in planned-BIRD3
power.bird3(es = 0.10, rho3 = .61, rho2 = 0, omega3 = .65, omega2 = 0,
            g3 = 1, r21 = 0.53, r2t3 = 0.04, r2t2 = 0,
            p = 0.50, n1 = 20, n2 = 5, n3 = 50)
```

We can also assume that a researcher substituted L2 parameters in the analyzed-msL3ig-BIRD2 for L2 parameters in planned-BIRD3. Thus, `rho3 = 0`, `omega3 = 0`, and `r2t3 = 0` as in the following.

```
# L2 parms for the analyzed-BIRD2 (L3 of the original BIRD3 ignored)
# can be substituted for L2 parms in planned-BIRD3
power.bird3(es = 0.10, rho2 = .61, rho3 = 0, omega2 = .65, omega3 = 0,
            g3 = 0, r21 = 0.53, r2t2 = 0.04, r2t3 = 0,
            p = 0.50, n1 = 20, n2 = 5, n3 = 50)
```

Results

Results presented in Table 3 answer *the "How do variance components shift when intermediate or top-level in a BIRD3 model is ignored?"* question. Table 3 presents unconditional variances for correctly specified BIRD3 and misspecified BIRD2 models. For correctly specified BIRD3 model, sources of variation in the outcome are attributed to L1 (students), L2 (classrooms), and L3 (schools) denoted as σ^2, τ_2^2, and τ_3^2, respectively. For the misspecified BIRD2 model, sources of variation in the outcome are attributed to L1 (students), and L2 (classrooms or schools) denoted as σ^2 and τ_2^2, respectively. Misspecified BIRD2 models could either ignore the intermediate level for which τ_2^2 refers to between-school variance or ignore top-level for which τ_2^2 refers to between classrooms variance. Table 3 demonstrates how variance parameters for an unconditional model shift when intermediate- or top-level was ignored. When the intermediate level was ignored in the BRID3 model, the L2 variance was distributed to the flanking levels in the new BIRD2 model. The variance distributed to the bottom level model was proportionally more (~80%) than the variance distributed to the top-level (~%20) in the new BIRD2 model. When the top-level was ignored, the bottom level remained the same; however, L2 variance in the new BIRD2 model was inflated approximately equal to the sum of L2 and L3 variance in the BIRD3 model. In both cases, the total variance was preserved.

Table 3
Unconditional Variance Parameters for BIRD3 and Misspecified BIRD2 Models

Analysis Model	Specification	Parameter	P1	P2	P3	P4	P5	P6	P7	P8
BIRD3	Correctly specified	σ^2	2.15	9.66	2.15	9.66	1.92	9.49	1.92	9.48
		τ_2^2	2.08	1.89	2.07	1.90	1.69	1.64	1.69	1.63
		τ_3^2	1.27	1.21	1.27	1.21	1.11	1.08	1.11	1.08
BIRD2	Intermediate level ignored in BIRD3	σ^2	3.83	11.18	3.83	11.19	3.29	10.81	3.29	10.80
		τ_2^2	1.66	1.58	1.67	1.58	1.44	1.39	1.43	1.39
BIRD2	Top-level ignored in BIRD3	σ^2	2.15	9.66	2.15	9.66	1.92	9.49	1.92	9.48
		τ_2^2	3.32	3.08	3.33	3.10	2.78	2.69	2.79	2.70

Note. The same symbols bear a different meaning in different models. σ^2 : L1 variance. τ_2^2 : L2 variance. τ_3^2 : L3 variance. Numbers in the table are averages of 5000 replications.

It is ideal for a researcher to analyze three-level data using the BIRD3 model. It is also desirable for a researcher to plan a BIRD3 model using parameters reported in existing scholarly work in which BIRD3 models were utilized. However, it is also possible for a researcher to analyze BIRD3 data using BIRD2 models where either intermediate level (classrooms) or top-level (schools) are ignored. Furthermore, it is also possible for a researcher to plan a BIRD3 model via using parameters from these misspecified BIRD2 models.

Results presented in Tables 4 to 7 answer "*How standard error of the treatment effect estimate (a L1 predictor) is affected by these misspecifications?*" question. When the intermediate level is ignored in a three-level BIRD3 model, it becomes a two-level BIRD2 model where the previous third level remains the top level. In addition to the shift in the variance components, which affects variance parameters in the new top and bottom levels, the sample size for the top-level remains the same (K). However, the sample size for L1 is now the combined sample size (nJ), whereas degrees of freedom for the test statistics do not change. MC simulations showed that power was slightly underestimated (see Table 4), whereas Type I error rates did not change substantially (see Table 5).

In contrast, when the top-level is ignored, the variance component shifts, and the sample size for the new top-level is now combined (JK). However, the sample size for the new bottom level remains the same (n), whereas degrees of freedom for the test statistics change due to the change in the number of top levels. As the top-level sample size is one of the most critical determinants of power, the change in top-level sample size alone was sufficient to overestimate power (see Table 6). However, Type I error rates were severely inflated (Table 7). Inflated Type I error offset the benefit of having an overpowered model.

The result of the MC simulation for the correctly specified BIRD3 model is provided in Tables 1A and 2A in Appendix A for comparison purposes. There was a close correspondence between MC-based power rates and those calculated via the `cosa` R package (see Table 1A). Type I error rates match 5% nominal rate (see Table 2A). This creates a baseline for further exploring power calculations in misspecified BIRD2 models.

Table 4
Comparison of Power Rates from Correctly Specified (BIRD3) and L2 Ignored (BIRD3) Model

Scenario	P1	P2	P3	P4	P5	P6	P7	P8
MC Power from BIRD3	0.44	0.30	0.74	0.52	0.45	0.26	0.72	0.45
MC Power from BIRD2	0.38	0.28	0.65	0.49	0.38	0.24	0.62	0.42
AD in Powers	-0.07	-0.03	-0.09	-0.03	-0.07	-0.01	-0.10	-0.03
RD in Powers	-15.09	-8.29	-12.05	-5.16	-16.02	-5.53	-13.57	-7.28

Note. AD: Absolute difference. RD: Relative difference (%). Power rates are based on 5000 replications.

Table 5
Comparison of Type I Error Rates from Correctly Specified (BIRD3) and L2 Ignored (BIRD2) Model

Scenario	T1	T2	T3	T4	T5	T6	T7	T8
MC Type I Error from BIRD3	0.06	0.06	0.05	0.05	0.05	0.06	0.05	0.05
MC Type I Error from BIRD2	0.05	0.06	0.06	0.05	0.06	0.06	0.05	0.05
AD in Type I Errors	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00
RD in Type I Errors	-5.90	2.14	8.95	-4.17	2.19	-2.45	-12.04	-8.65

Note. AD: Absolute difference. RD: Relative difference (%). Type I error rates are based on 5000 replications.

Table 6
Comparison of Power Rates from Correctly Specified (BIRD3) and L3 Ignored (BIRD2) Model

Scenario	P1	P2	P3	P4	P5	P6	P7	P8
MC Power from BIRD3	0.44	0.30	0.74	0.52	0.45	0.26	0.72	0.45
MC Power from BIRD2	0.62	0.43	0.86	0.66	0.63	0.34	0.84	0.54
AD in Powers	0.18	0.13	0.12	0.14	0.19	0.08	0.12	0.09
RD in Powers	40.68	43.44	16.00	26.01	41.50	30.66	16.52	20.36

Note. AD: Absolute difference. RD: Relative difference (%). Power rates are based on 5000 replications.

Table 7
Comparison of Type I Error Rates from Correctly Specified (BIRD3) and L3 Ignored (BIRD) Model

Analysis Model	T1	T2	T3	T4	T5	T6	T7	T8
MC Type I Error from BIRD3	0.06	0.06	0.05	0.05	0.05	0.06	0.05	0.05
MC Type I Error from BIRD2	0.15	0.15	0.16	0.14	0.14	0.10	0.14	0.10
AD in Type I Errors	0.09	0.09	0.11	0.09	0.09	0.05	0.09	0.05
RD in Type I Errors	159.03	158.93	208.56	168.94	161.68	79.72	156.20	90.60

Note. AD: Absolute difference. RD: Relative difference (%). Type I error rates are based on 5000 replications.

The results presented earlier were related to the analysis phase, and all numbers reported therein was based on MC simulation. One possibility mentioned earlier was to use parameters from a misspecified BIRD2 model to design a BIRD2 or BIRD3 study. One could obtain parameters needed in power calculations via analyzing three-level data using the BIRD2 model (assuming either intermediate/top-level is not available or ignored) or using parameters reported in scholarly work in which the BIRD2 model was utilized.

Results presented in Tables 8 and 9 answers "*Can we use design parameters from misspecified BIRD2 model (either intermediate or top-level ignored) for a prospective BIRD2 or BIRD3 design?*" question. Table 8 presents the misspecified BIRD2 model where the intermediate level was ignored. Power rates for a prospective study were calculated considering three cases; parameters obtained from the misspecified BIRD2 analysis can be used (i) for planning a BIRD2 design, (ii) for planning a BIRD3 design where parameters of L2 in BIRD2 analysis were substituted for parameters of L3 in BIRD3 design (thus parameters of L2 in BIRD3 design were all constrained to zero), and (iii) for planning a BIRD3 design where parameters of L2 in BIRD2 analysis were substituted for parameters of L2 in BIRD3 design (thus parameters of L3 in BIRD3 design are all constrained to zero).

Table 8
Power Rates for the Misspecified BIRD2 Model (L2 Ignored)

Scenario	P1	P2	P3	P4	P5	P6	P7	P8
$\hat{\xi}_{100}$	0.24	0.25	0.25	0.25	0.25	0.24	0.25	0.25
$SE(\hat{\xi}_{100})$	0.15	0.18	0.11	0.13	0.15	0.20	0.11	0.14
$ES(\hat{\xi}_{100})$	0.10	0.07	0.11	0.07	0.12	0.07	0.11	0.07
ρ_2	0.30	0.12	0.30	0.12	0.30	0.11	0.30	0.11
ω_2	0.54	0.49	0.53	0.48	0.66	0.57	0.65	0.57
R_1^2	0.22	0.04	0.22	0.04	0.22	0.03	0.22	0.03
R_{2T}^2	0.08	0.07	0.07	0.06	0.07	0.07	0.07	0.06
p	0.50	0.50	0.50	0.50	0.20	0.20	0.20	0.20
ρ_{TS}	0.80	0.80	0.80	0.80	0.70	0.70	0.70	0.70
K	50	50	100	100	50	50	100	100
$SE_{MC}(\hat{\xi}_{100})$	0.15	0.19	0.11	0.13	0.15	0.21	0.11	0.14
MC Power	0.38	0.28	0.65	0.49	0.38	0.24	0.62	0.42
(i) cosa R Package (Plan BIRD2)	0.33	0.24	0.67	0.44	0.38	0.22	0.59	0.40
(ii) cosa R Package (Plan BIRD3: L2 Parms = 0)	0.33	0.24	0.67	0.44	0.38	0.22	0.59	0.40
(iii) cosa R Package (Plan BIRD3: L3 Parms = 0)	0.64	0.33	0.95	0.58	0.73	0.29	0.92	0.53

Note. Results are based on 5000 replications. $\hat{\xi}_{100}$: Treatment effect. SE: Standard Error. ES: Effect size. ρ_2: Proportion of variance in the outcome between L2 units. ω_2: Treatment effect heterogeneity across L2 units. R_1^2: Proportion of variance in the outcome explained L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained L2 covariates. *p*: Proportion of subjects fall below (or above) cutoff score on the assignment variable. ρ_{TS}: Correlation between the assignment variable and the treatment status. *nJ*: Average number of L1 units per L2 units, set to 100. *K*: Number of L3 units.

Table 9
Power Rates for the Misspecified BIRD2 Model (L3 Ignored)

Scenario	P1	P2	P3	P4	P5	P6	P7	P8
$\hat{\xi}_{100}$	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
$SE(\hat{\xi}_{100})$	0.10	0.14	0.07	0.10	0.11	0.17	0.07	0.12
$ES(\hat{\xi}_{100})$	0.10	0.07	0.11	0.07	0.12	0.07	0.11	0.07
ρ_2	0.61	0.24	0.61	0.24	0.59	0.22	0.59	0.22
ω_2	0.65	0.52	0.65	0.52	0.77	0.60	0.77	0.59
R_1^2	0.53	0.07	0.54	0.07	0.48	0.05	0.48	0.05
R_{2T}^2	0.04	0.05	0.04	0.04	0.04	0.04	0.03	0.04
p	0.50	0.50	0.50	0.50	0.20	0.20	0.20	0.20
ρ_{TS}	0.80	0.80	0.80	0.80	0.70	0.70	0.70	0.70
JK	250	250	500	500	250	250	500	500
$SE_{MC}(\hat{\xi}_{100})$	0.14	0.19	0.10	0.13	0.14	0.20	0.10	0.14
MC Power	0.62	0.43	0.86	0.66	0.63	0.34	0.84	0.54
(i) cosa R Package (Plan BIRD2)	0.62	0.34	0.94	0.59	0.71	0.30	0.90	0.54
(ii) cosa R Package (Plan BIRD3: L2 Parms = 0)	0.20	0.20	0.23	0.19	0.23	0.18	0.20	0.18
(iii) cosa R Package (Plan BIRD3: L3 Parms = 0)	0.61	0.33	0.69	0.33	0.70	0.30	0.62	0.30

Note. Results are based on 5000 replications. $\hat{\xi}_{100}$: Treatment effect. SE: Standard Error. ES: Effect size. ρ_2: Proportion of variance in the outcome between L2 units. ω_2: Treatment effect heterogeneity across L2 units. R_1^2: Proportion of variance in the outcome explained L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained L2 covariates. *p*: Proportion of subjects fall below (or above) cutoff score on the assignment variable. ρ_{TS}: Correlation between the assignment variable and the treatment status. *n*: Average number of L1 units per L2 units, set to 20. *JK*: Number of L2 units. AD: Absolute difference. RD: Relative difference.

For cases (i) and (ii) in Table 8 (L2 ignored), calculated power rates slightly underestimated MC-based power rates for misspecified BIRD2 and underestimated MC-based power rates for correctly specified BIRD3. However, in case (iii), calculated power rates were somewhat optimistic, substantially exceeding MC-based power rates of both designs. On the contrary, Table 9 (L3 ignored) portray a different picture. Calculated power rates were unstable for all cases. Calculated power rates in case (i) were underestimated or overestimated compared to MC-based power rates for misspecified BIRD2 and overestimated compared to MC-based power rates for correctly specified BIRD3, in (ii) they

were severely underpowered compared to both, and in (iii) they were unstable considering both. The term "unstable" means we observed no trend regarding the magnitude or direction of the difference from MC-based power rates.

Discussion

From analysis perspective, when intermediate-level was ignored in BIRD3, the majority of the variance in the ignored level shifts to the new bottom level, and a small portion of the variance shifts to the new top-level in BIRD2. These results are in line with findings in Moerbeek (2004), van Den Noortgate et al. (2005), and Opdenakker and van Damme (2000). This shift in variance components causes a slight underestimation of power rates. It can be neglected if the variance at the intermediate level is small to moderate, to begin with, confirming findings in Zhu et al. (2011). However, classroom-level variance can exceed school-level variance in practice (Goldstein, 2011; Muthen, 1991). Even if sample sizes at the intermediate level are very small, this problem can be addressed by using bootstrapping or Bayesian methods (Goldstein, 2011) or introducing L2 information as fixed effects into the model (van Den Noorthgate et al., 2005). Another way to decide whether to acknowledge or ignore an intermediate level is to base the modeling decision on the model fit (Opdenakker & Van Damme, 2000). Suppose the chi-square test of difference is meaningful between the model that ignores and the model that acknowledges the intermediate level. In that case, it is advisable to acknowledge the intermediate level and pursue the analysis accordingly. If the data permits, to mitigate the problem, the least an analyst could do is to introduce predictors belonging to the ignored level (Opdenakker & Van Damme, 2000). However, the reason to ignore a level is the absence of information on that level. If any information is available (e.g., covariates), the L2 or L3 membership can be constructed.

These remedies might apply to the analysis phase but are not necessarily needed for the planning phase. Considering Type I errors did not change substantially, one could use parameters from the BIRD2 model to design a BIRD2 or BIRD3 model. The study will be adequately powered during the analysis phase as long as the design satisfies the desired level of power rate, except when L2 parameters of a misspecified BIRD2 model is substituted for L2 parameters in a future BIRD3 design. In this case, the test statistics will be underpowered in the analysis phase. The top-level sample size could be oversampled to make up for this, but to what extent it should be inflated is unknown ahead of the study.

Ignoring the top level is more problematic, although the variance component at the third level is negligible. When the top-level is ignored, the variance of the ignored level in BIRD3 shifts to the new top-level in BIRD2, in line with findings in Moerbeek (2004) and van Den Noortgate et al. (2005). This shift in variance component and increased sample size at L2 causes overestimation of power rates. Compared to the ignoring intermediate-level, the distortion in ignoring the top-level is more pronounced as the top-level sample size change dramatically. On its own, this would not constitute a significant problem if the top-level sample size is inflated during the planning phase. However, as mentioned earlier, to what extent it should be inflated is unknown ahead of the study. Regardless, it should be avoided at all costs because Type I error rates were severely inflated.

Limitations

Results and their implications are limited to the simulated scenarios. When the variance component for the intermediate level is significant, results may differ. Furthermore, ignoring a level may also mean omitting relevant variables at that level. This means ignoring a level also comes with an omitted variable bias, which complicates misspecification. Functional form misspecification is another topic that deserves attention. Bulus (2022) recently found that for balanced RDD designs ($p = 0.50$), power rates for a linear form of the score variable, linear form interacting with the treatment variable, or quadratic form of the score variable does not change. However, a quadratic form of the score variable interacting with the treatment variable requires a larger sample size to reach the same power rate of lower polynomial forms. He also found that power rates may differ across different functional form specifications for unbalanced designs (e.g., $p = 0.20$). In this study, only the linear form of the score variable was considered. The incorrect functional form may complicate misspecification even further.

References

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). Evaluation of response to intervention practices for elementary school reading. NCEE 2016-4000. *National Center for Education Evaluation and Regional Assistance*. https://files.eric.ed.gov/fulltext/ED560820.pdf

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* Guilford Press.

Bulus, M. (2022). Minimum detectable effect size computations for cluster-level regression discontinuity: Specifications beyond the linear functional form. *Journal of Research on Education Effectiveness, 15*(1), 151-177. https://doi.org/10.1080/19345747.2021.1947425

Bulus, M., & Dong, N. (2021a). Bound constrained optimization of sample sizes subject to monetary restrictions in planning of multilevel randomized trials and regression discontinuity studies. The *Journal of Experimental Education*, *89*(2), 379–401. https://doi.org/10.1080/00220973.2019.1636197

Bulus, M., & Dong, N. (2021b). cosa: Bound constrained optimal sample size allocation. R package version 2.1.0. https://CRAN.R-project.org/package=cosa

Bulus, M., & Dong, N. (2022). Minimum detectable effect size computations for blocked individual-level regression discontinuity: Specifications beyond the linear functional form. *Manuscript in preparation.*

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment long-run impacts of double-dose algebra. *Journal of Human Resources*, *50*(1), 108-158. https://doi.org/10.3368/jhr.50.1.108

Deke, J., Dragoset, L., Bogen, K., & Gill, B. (2012). Impacts of Title I Supplemental Educational Services on student achievement. NCEE 2012-4053. *National Center for Education Evaluation and Regional Assistance*. https://ies.ed.gov/ncee/pubs/20124053/pdf/20124053.pdf

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

Finch, W. H., & Bolin, J. E. (2017). *Multilevel modeling using Mplus*. CRC Press.

Fox, J. (1997). A*pplied regression analysis, linear models, and related methods*. Sage Publications.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922): John Wiley & Sons.

Harrington, J. R., Muñoz, J., Curs, B. R., & Ehlert, M. (2016). Examining the impact of a highly targeted state-administered merit aid program on brain drain: Evidence from a regression discontinuity analysis of Missouri's Bright Flight program. *Research in Higher Education*, *57*(4), 423-447. https://doi.org/10.1007/s11162-015-9392-9

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Routledge.

Hustedt, J. T., Jung, K., Barnett, W. S., & Williams, T. (2015). Kindergarten readiness impacts of the Arkansas Better Chance State Prekindergarten Initiative. *The Elementary School Journal*, *116*(2), 198-216. https://doi.org/10.1086/684105

Jenkins, J. M., Farkas, G., Duncan, G. J., Burchinal, M., & Vandell, D. L. (2016). Head Start at ages 3 and 4 versus Head Start followed by state Pre-K which is more effective? *Educational evaluation and policy analysis*, *38*(1), 88-112. https://doi.org/10.3102%2F0162373715587965

Klerman, J. A., Olsho, L. E., & Bartlett, S. (2015). Regression discontinuity in prospective evaluations: The case of the FFVP evaluation. *American Journal of Evaluation*, *36(*3), 403-416. https://doi.org/10.1177%2F1098214014553786

Konstantopoulos, S., & Shen, T. (2016). Class size effects on mathematics achievement in Cyprus: evidence from TIMSS. *Educational Research and Evaluation,* *22*(1-2), 86-109. https://doi.org/10.1080/13803611.2016.1193030

Konu, A. I., Lintonen, T. P., & Autio, V. J. (2002). Evaluation of well-being in schools–a multilevel analysis of general subjective well-being. *School Effectiveness and School Improvement,* *13*(2), 187-200.

https://doi.org/10.1076/sesi.13.2.187.3432

Leeds, D. M., McFarlin, I., & Daugherty, L. (2017). Does student effort respond to incentives? Evidence from a guaranteed college admissions program. *Research in Higher Education*, *58*(3), 231-243. http://dx.doi.org/10.1007/s11162-016-9427-x

Ludwig, J., & Miller, D. L. (2005). *Does Head Start improve children's life chances? Evidence from a regression discontinuity design* (No. w11702). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w11702/w11702.pdf

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education, 32*(3), 397-429. https://doi.org/10.1080/03054980600776589

Luyten, H., Peschar, J., & Coe, R. (2008). Effects of schooling on reading performance, reading engagement, and reading activities of 15-year-olds in England. *American Educational Research Journal*, *45*(2), 319-342. https://doi.org/10.3102%2F0002831207313345

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, *142*(2), 829-850. https://doi.org/10.1016/j.jeconom.2007.05.015

Manatunga, A. K., Hudgens, M. G., & Chen, S. (2001). Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *43*(1), 75-86. http://dx.doi.org/10.1002/1521-4036(200102)43:1%3C75::AID-BIMJ75%3E3.0.CO;2-N

May, H., Sirinides, P. M., Gray, A., & Goldsworthy, H. (2016). Reading recovery: An evaluation of the four-year i3 scale-up. https://files.eric.ed.gov/fulltext/ED593261.pdf

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149. https://doi.org/10.1207/s15327906mbr3901_5

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338-354. http://www.jstor.org/stable/1434897

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257. https://doi.org/10.3102%2F01623737026003237

Opdenakker, M.-C., & van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement, 11*(1), 103-130. https://doi.org/10.1076/0924-3453(200003)11:1;1-A;FT103

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17. https://doi.org/10.2307/2112482

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

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course-taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis*, *32*(4), 498-520. https://doi.org/10.3102%2F0162373710382655

Schochet, P. Z. (2008). Technical methods report: Statistical power for regression discontinuity designs in education evaluations. NCEE 2008-4026. *National Center for Education Evaluation and Regional Assistance*. https://files.eric.ed.gov/fulltext/ED511782.pdf

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*,*34*(2), 238-266. https://doi.org/10.3102%2F1076998609332748

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications.

Singer, J. (1987). An intraclass correlation for analyzing multilevel data. *Journal of Experimental Education*, *55*(4), 219–228. https://doi.org/10.1080/00220973.1987.10806457

van Den Noortgate, W., Opdenakker, M.-C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement, 16*(3), 281-303. https://doi.org/10.1080/09243450500114850

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy analysis and management*, *27*(1), 122-154. https://doi.org/10.1002/pam.20310

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2011). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis*, *34*(1), 45-68. https://doi.org/10.3102%2F0162373711423786

Appendix

Table 1A
Power Rates for Correctly Specified BIRD3 Design

Scenario	P1	P2	P3	P4	P5	P6	P7	P8
$\hat{\xi}_{100}$	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
$SE(\hat{\xi}_{100})$	0.14	0.18	0.10	0.13	0.14	0.19	0.10	0.14
$ES(\hat{\xi}_{100})$	0.10	0.07	0.11	0.07	0.12	0.07	0.11	0.07
ρ_2	0.38	0.15	0.38	0.15	0.36	0.13	0.36	0.13
ρ_3	0.23	0.09	0.23	0.09	0.23	0.09	0.23	0.09
ω_2	0.77	0.57	0.77	0.56	0.90	0.64	0.91	0.65
ω_3	0.47	0.47	0.46	0.46	0.54	0.52	0.52	0.52
R_1^2	0.53	0.07	0.54	0.07	0.48	0.05	0.48	0.05
R_{T2}^2	0.06	0.07	0.06	0.06	0.05	0.07	0.05	0.06
R_{T3}^2	0.13	0.11	0.11	0.09	0.13	0.14	0.11	0.09
p	0.50	0.50	0.50	0.50	0.20	0.20	0.20	0.20
ρ_{TS}	0.80	0.80	0.80	0.80	0.70	0.70	0.70	0.70
K	50	50	100	100	50	50	100	100
$SE_{MC}(\hat{\xi}_{100})$	0.14	0.18	0.10	0.13	0.14	0.20	0.10	0.14
MC Power	0.44	0.30	0.74	0.52	0.45	0.26	0.72	0.45

Note. Results are based on 5000 replications. $\hat{\xi}_{100}$: Treatment effect. SE: Standard Error. ES: Effect size. ρ_2: Proportion of variance in the outcome between L2 units. ρ_3: Proportion of variance in the outcome between L3 units. ω_2: Treatment effect heterogeneity across L2 units. ω_3: Treatment effect heterogeneity across L3 units. R_1^2: Proportion of variance in the outcome explained L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained L2 covariates. R_{T3}^2: Proportion of variance in the treatment effect explained L3 covariates. *p*: Proportion of subjects fall below (or above) cutoff score on the assignment variable. ρ_{TS}: Correlation between the assignment variable and the treatment status. *n*: Average number of L1 units per L2 units, which is set to 20. *J*: Average number of L2 units per L3 units, which is set to 5. *K*: Number of L3 units.

Table 2A
Type I Error Rates for Correctly Specified BIRD3 Design

Scenario	T1	T2	T3	T4	T5	T6	T7	T8
$\hat{\xi}_{100}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
$SE(\hat{\xi}_{100})$	0.14	0.18	0.10	0.13	0.14	0.19	0.10	0.14
$ES(\hat{\xi}_{100})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ρ_2	0.39	0.15	0.39	0.15	0.36	0.13	0.36	0.13
ρ_3	0.23	0.10	0.23	0.10	0.24	0.09	0.24	0.09
ω_2	0.77	0.57	0.77	0.56	0.90	0.64	0.91	0.65
ω_3	0.47	0.47	0.46	0.46	0.54	0.52	0.53	0.52
R_1^2	0.51	0.06	0.51	0.06	0.46	0.05	0.46	0.05
R_{T2}^2	0.06	0.07	0.06	0.06	0.05	0.07	0.05	0.06
R_{T3}^2	0.13	0.10	0.11	0.09	0.13	0.13	0.12	0.10
p	0.50	0.50	0.50	0.50	0.20	0.20	0.20	0.20
ρ_{TS}	0.80	0.80	0.80	0.80	0.70	0.70	0.70	0.70
K	50	50	100	100	50	50	100	100
$SE_{MC}(\hat{\xi}_{100})$	0.14	0.18	0.10	0.13	0.14	0.20	0.10	0.14
MC Type I Error	0.06	0.06	0.05	0.05	0.05	0.06	0.05	0.05

Note. Results are based on 5000 replications. $\hat{\xi}_{100}$: Treatment effect. SE: Standard Error. ES: Effect size. ρ_2: Proportion of variance in the outcome between L2 units. ρ_3: Proportion of variance in the outcome between L3 units. ω_2: Treatment effect heterogeneity across L2 units. ω_3: Treatment effect heterogeneity across L3 units. R_1^2: Proportion of variance in the outcome explained L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained L2 covariates. R_{T3}^2: Proportion of variance in the treatment effect explained L3 covariates. *p*: Proportion of subjects fall below (or above) cutoff score on the assignment variable. ρ_{TS}: Correlation between the assignment variable and the treatment status. *n*: Average number of L1 units per L2 units, which is set to 20. *J*: Average number of L2 units per L3 units, which is set to 5. *K*: Number of L3 units.