Article History
Received: 06.02.2022
Received in revised form: 04.06.2022
Accepted: 04.06.2022
Available online: 30.06.2022
Article Type: Research Article

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[1], Nianbo Dong[2]
[1]Adıyaman University, Adıyaman, Türkiye
[2] University of North Carolina at Chapel Hill, North Carolina, USA

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 Univesity Journal of Educational Sciences, 12*(1), 42-55.

Adiyaman University Journal of Educational Sciences

Volume 12, Number 1, June 2022, Page 42-55 ISSN: 2149-2727
DOI: http://dx.doi.org/10.17984/adyuebd.1068923

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**
Adıyaman University, Adıyaman, Türkiye

Nianbo Dong
University of North Carolina at Chapel Hill, North Carolina, USA

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 nested in classrooms/teachers nested in schools), whether unwittingly during data analysis or due to resource constraints during the planning phase. This study investigates the consequences of ignoring intermediate or top level in blocked three-level regression discontinuity designs (BIRD3; treatment is at level 1) during data analysis and planning. Monte Carlo simulation results indicated that ignoring a level during analysis did not affect the accuracy of treatment effect estimates; however, it affected the precision (standard errors, power, and Type I error rates). Ignoring the 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 overestimated power rates; however, severe inflation in Type I error deemed this strategy ineffective. As for the design phase, when the intermediate level was ignored, it is viable to use parameters from a two-level blocked regression discontinuity model (BIRD2) to plan a BIRD3 design. However, level 2 parameters from the BIRD2 model should be substituted for level 3 parameters in the BIRD3 design. When the top level was ignored, using parameters from the BIRD2 model to plan a BIRD3 design should be avoided.

Keywords: blocked regression discontinuity designs, multilevel models, hierarchical linear models, ignoring a level of nesting, power analysis

Introduction

One fundamental assumption of Ordinary Least Squares (OLS) regression is that observations are conditionally independent. This assumption is violated when errors are not independent of each other (presenting autocorrelation) due to the 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 estimates 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) is a compelling option for remedying the violation of independent errors when the 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 (as predictors). One can also study the influence of contextual factors on the estimates of predictors (as moderators). 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 substantive research questions owing to rapid advances in its 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).

* This article was produced from the corresponding author's doctoral dissertation titled "Design Considerations in Three-level Regression Discontinuity Studies" from the 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*

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 a lack of economic resources, missing administrative records, or researchers' unwitting ignorance. 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 to (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 level of nesting occurs when classroom-level information is ignored. However, variance attributed to the classroom level can exceed that of the 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 the 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 indicated that regression discontinuity designs (RDDs) are not exempt from this practice (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 and standard errors (Moerbeek, 2004; Opdenakker & Van Damme, 2000; Van den Noortgate et al., 2005; Zhu et al., 2011). From this point forward, 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 inflation in the L2 variance was approximately equal to the ignored amount at L3. 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 components at L2 and L1. However, the ignored variance at L4 was transferred to the variance at L3.

Ignoring an intermediate level is 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 inflated variance components at the flanking levels. For example, the variance was distributed to L2 and L4 when L3 was ignored. This finding is in line with 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 the predictor variable was measured, and sample sizes at one or more levels.

Effects of Ignoring a Level of Nesting on Standard Errors

The literature has already established that fixed effect estimates 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 the variance component of a given level is affected due to ignoring a level of nesting, standard errors of the estimates are also 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 the ignored level, and sample sizes at each level. For example, ignoring L2 inflates standard errors for fixed effect estimates at L1, resulting in a loss of power but not those at L3 (Moerbek, 2004). However, as Moerbek (2004) noted, if the proportion of variance 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 empirical 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 adjacent levels were affected. When level 4 was ignored, the standard error of the estimate for predictors at L3

[†] A balanced design means having the same number of lower-level units per higher-level unit. For example, a balanced two-level design would have *n* number of level 1 units for each level 2 unit.

was affected. When L3 was ignored in balanced data, the standard error of the estimate for predictors at L2 increased. When the data was unbalanced and L3 was ignored, the standard error of the estimates for predictors at level 4 decreased.

Opdenakker and Van Damme (2000) found that regardless of which level is ignored, the standard error of the intercept was underestimated. However, when L4 was ignored, the standard error of the estimates at L1 and L2 was not affected as much. Zhu et al. (2011) extended previous work on ignoring a level of nesting by mainly focusing on the design phase of cluster-randomized trials rather than analysis, although results apply to both. In particular, the 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 or standard error of the estimate for predictors at the top level. Additionally, they concluded that using design parameters from a two-level model to design a three-level study did not pose a substantial threat to the precision of the treatment effect at the top level.

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

From 2000 onward, several studies used RDD with a discontinuity at L1. These studies, one way or another, adjusted their estimates for clustering. About a quarter of them used the MLM framework to adjust for clustering (e.g., Hustedt et al., 2015; Luyten, 2006; Luyten et al., 2008; May et al., 2016), and about a quarter of them used Lee and Card (2008) method (e.g., 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 (e.g., 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 the individual level cutoff 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 a single-level RDD, district-level information was included in the model as fixed effects. 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. Similarly, 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. Luyten (2006) and Luyten et al. (2008) analyzed the data using a two-level RDD model where the effect of an extra year of schooling was assumed to vary across schools randomly. 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 model where the program effect was assumed to vary across schools randomly. In summary, four RDDs relying on individual level cutoff-based assignment and the MLM framework could have been analyzed by acknowledging the classroom level information (intermediate level) or district or state-level fixed effects (top-level).

Problem Statement

Drawing from four multi-site empirical elementary and secondary school datasets, Zhu et al. (2011) concluded that ignoring the intermediate level did not pose a substantial threat to the design and analysis of three-level cluster-randomized trials. 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 analysis phase. From a design perspective, Zhu et al. (2011) showed that using design parameters from a two-level model (the intermediate level was ignored) is viable for designing a three-level study where the treatment variable is at the top level. Whether these findings can be extended to designs with the L1 treatment variable is unclear. In this study, within the context of blocked three-level RDD design (BIRD3), we investigate whether it is plausible to use parameters from a misspecified blocked two-level RDD design (BIRD2) model (either intermediate or top-level in BIRD3 design is ignored) to plan 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 is the precision of the treatment effect estimate (an L1 predictor) affected by these misspecifications?
3. Can we use design parameters from a misspecified BIRD2 model (where intermediate or top level in BIRD3 design was ignored) to plan a future BIRD3 design?

Method

Consider a sample with three levels of nesting structure (e.g., students as L1 – classrooms as L2 – schools as 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 intercept and treatment effect is random across L2 and L3 units. Also, assume that the data is balanced: *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, the power rate of the average treatment effect approximates formula-based power rates in the `cosa` R package (Bulus & Dong, 2021a; Bulus & Dong, 2021b) and PowerUp! software (Dong & Maynard, 2013).

Next, we describe statistical models (unconditional, treatment-only, and full models) for the correctly specified BIRD3 model. We also define standardized variance parameters such as intra-class correlation coefficients, R-squared values, and treatment effect heterogeneity. These standardized parameters can be used in subsequent precision calculations (statistical power, minimum required sample size, and minimum detectable effect size). Furthermore, we also describe standardized standard error formulas for the L1 treatment effect in BIRD3 and BIRD2 designs. Standardized standard error formulas require standardized variance parameters as input and provide the basis for precision calculations.

Statistical Models

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

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

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

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

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 $\begin{pmatrix}\varsigma_{00\mathrm{k}}\\ \varsigma_{10\mathrm{k}}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_{3|T}^2 & \tau_{3T3}\\ \tau_{3T3} & \tau_{T3}^2\end{pmatrix}\right)$.

Full Model

The following model is used to generate data for Monte Carlo simulations. It is also used to obtain variance parameters $\sigma_{|T,S,X}^2$, $\tau_{2|W}^2$, and $\tau_{3|V}^2$, as defined below, which are used to calculate various standardized parameters along with the parameters from the unconditional and treatment-only model. This model also estimates Monte Carlo-based treatment effect, standard error, and power and Type I error rates for a given scenario.

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

$$\gamma_{10k} = \xi_{100} + \xi_{101}V_k + \varsigma_{10k}$$
$$\gamma_{20k} = \xi_{200}$$
$$\gamma_{30k} = \xi_{300}$$
$$\gamma_{01k} = \xi_{010}$$
$$\gamma_{11k} = \xi_{110},$$

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

Using parameters from unconditional, treatment only, and full models define

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

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

$\omega_2 = \frac{\tau_{T2}^2}{\tau_2^2}$, which is the treatment effect heterogeneity across L2 units;

$\omega_3 = \frac{\tau_{T3}^2}{\tau_3^2}$, which is the treatment effect heterogeneity across L3 units;

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

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

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

Next, we provide standardized standard error formulas for treatment effect in BIRD3 and BIRD2 design which are re-parameterized using standardized variance parameters defined above.

Standardized Standard Error for the Correctly Specified BIRD3 Model

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

$$SE\left(\hat{\xi}_{100}\right) = \sqrt{\frac{\omega_3\rho_3(1-R^2_{T3})}{K} + \frac{\omega_2\rho_2(1-R^2_{T2})}{KJ} + \frac{(1-\rho_3-\rho_2)(1-R^2_1)(RDDE)}{KJnp(1-p)}}$$

where RDDE is the regression discontinuity design effect and takes the form of $RDDE = 1/(1-\rho^2_{TS})$ when the linear form of the score variable is considered (Bulus, 2022; Bulus & Dong, 2022; Schochet, 2008, 2009). ρ^2_{TS} is the squared correlation between treatment and score variables. It is 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).

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: Scenario 1 yields $\rho_2 \cong 0.40$ and $\rho_3 \cong 0.20$, and Scenario 2 yields $\rho_2 \cong 0.15$ and $\rho_3 \cong 0.10$ approximately.

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 $n = 20$ and $J = 5$ constant across all the scenarios. Sample sizes were chosen to approximate those commonly encountered in education. Although $J = 5$ may not be as common, it is an ideal minimum number to obtain consistent variance estimates. In total, there were eight scenarios for statistical power simulation (P1-P8) and eight scenarios for Type I error simulation (T1-T8).

Analysis

We used PROC MIXED in SAS with default restricted maximum likelihood (REML) estimation and unstructured (UN) variance-covariance structure. 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 the "Null Model," "Treatment-only Model," and "Full Model." 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 are based on averages of 5000 replications. There were 5000 rows for estimates, standard errors, and variance parameters, but only their averages were used to obtain standardized variance parameters.

Power Calculations

Averages of 5000 raw estimates were transformed into standardized parameters according to definitions in the "Null Model," "Treatment-only Model," and "Full Model" described in the earlier section. Then, the standardized parameters were used in `power.bird3()` function in the `cosa` R library (Bulus & Dong, 2021a, 2021b). Model parameters, corresponding arguments, and their possible range are defined in Table 1. There are four combination of power calculations. One could ignore either intermediate or top level in a BIRD3 design, and use L2 parameters obtained from a BIRD2 model for either L2 or L3 parameters in a BIRD3 design.

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	$R_1^2 = 1 - \frac{\sigma_{\lvert T,S,X}^2}{\sigma^2}$	$R_{T2}^2 = 1 - \frac{\tau_{T2\lvert W}^2}{\tau_{T2}^2}$	$R_{T3}^2 = 1 - \frac{\tau_{T3\lvert V}^2}{\tau_{T3}^2}$	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	$n_1 \in N^+$	$n_2 \in N^+$	$n_3 \in N^+$		

When intermediate level is ignored, and L2 parameters from a BIRD2 model is used for L3 parameters in a future BIRD3 design the following code format is used. Note that L2 parameters in a future BIRD3 design are constrained to zero; thus, `rho2 = 0`, `omega2 = 0`, and `r2t2 = 0`.

```
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)
```

When intermediate level is ignored, and L2 parameters from a BIRD2 model is used for L2 parameters in a future BIRD3 design the following code format is used. Note that L3 parameters in a future BIRD3 design are constrained to zero; thus, `rho3 = 0`, `omega3 = 0`, and `r2t3 = 0`.

```
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)
```

When top level is ignored, and L2 parameters from a BIRD2 model is used for L3 parameters in a future BIRD3 design the following code format is used. Note that L2 parameters in a future BIRD3 design are constrained to zero; thus, `rho2 = 0`, `omega2 = 0`, and `r2t2 = 0`.

```
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)
```

When top level is ignored, and L2 parameters from a BIRD2 model is used for L2 parameters in a future BIRD3 design the following code format is used. Note that L3 parameters in a future BIRD3 design are constrained to zero; thus, `rho3 = 0`, `omega3 = 0`, and `r2t3 = 0` as in the following.

```
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 2 answer the *"How do variance components shift when intermediate or top level in a BIRD3 model is ignored?"* question. Table 2 presents unconditional variances for correctly specified BIRD3 and misspecified BIRD2 models. For the 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. In the misspecified BIRD2 models, one could either ignore the intermediate level for which τ_2^2 refers to the between-school variance or the top level for which τ_2^2 refers to the between-classroom variance. In what follows, we use the term "model" to refer to the analysis model and "design" to refer to the planned model. For example, the "BIRD3 model" refers to the analysis model, whereas the "BIRD3 design" refers to the planned model.

Table 2 demonstrates how variance parameters for the unconditional model shift when intermediate or top level was ignored. The variance of the ignored level was distributed to the flanking levels when the intermediate level was ignored. The variance distributed to the bottom level model was proportionally more (~80%) than the variance distributed to the top level (~%20). The variance of the bottom level remained the same when the top level was ignored. However, variance of the top level in the new BIRD2 model was approximately equal to the sum of L2 and L3 variance in the BIRD3 model. In both cases, the total variance was preserved.

Table 2. 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 is ignored	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 is ignored	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

The same symbol bears different meanings in different models. σ^2: L1 variance. τ_2^2: L2 variance. τ_3^2: L3 variance. Numbers are averages of 5000 replications.

It is ideal for a researcher to analyze data with three levels of nesting using the BIRD3 model. It is also desirable for a researcher to plan a BIRD3 design using parameters reported in existing scholarly work in which BIRD3 models were utilized. However, it is also possible for a researcher to analyze data with three levels of nesting using the BIRD2 model where either intermediate level (classrooms) or top-level (schools) is ignored. Results presented in Tables 3 to 6 answer the "*How is the precision of the treatment effect estimate (an L1 predictor) affected by these misspecifications?*" question. When the intermediate level is ignored in a BIRD3 model, it becomes a BIRD2 model where the previous third level remains the top level. The variance component of the ignored level is distributed to the new top and bottom levels. The sample size for the top level remains the same (K); however, the sample size for the bottom level is now the combined sample size (nJ). Finally, the degrees of freedom for the test statistic does not change. When the top-level is ignored; however, the variance component at the bottom level does not change, whereas the variance of the ignored level is conveyed to the new top level. The sample size for the new top level is now combined (JK), whereas the sample size for the new bottom level remains the same (n). On the contrary, the degrees of freedom for the test statistic changes due to the increased top-level sample size.

When the intermediate level was ignored, MC simulation results indicated that power rates were slightly underestimated (see Table 3), whereas Type I error rates did not change substantially (see Table 4). In contrast, when the top-level was ignored, power rates were overestimated, and Type I errors were severely inflated. As the top-level sample size is one of the most critical determinants of power, the change in the top-level sample size alone was sufficient to overestimate power (see Table 5). However, Type I error rates were severely inflated (see Table 6). Inflated Type I error rates 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 the 5% nominal rate (see Table 2A). The tables in the Appendix A provide a baseline for further exploring and comparing power calculations in the following sections.

Table 3. Comparison of power rates from BIRD3 and L2-ignored 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

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

Table 4. Comparison of Type I error rates from BIRD3 and L2-ignored 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

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

Table 5. Comparison of power rates from BIRD3 and L3-ignored 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

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

Table 6. Comparison of Type I error rates from BIRD3 and L3-ignored 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

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. A researcher can use parameters from a misspecified BIRD2 model (assuming intermediate/top level is not available or ignored) to plan a BIRD3 design. Results presented in Tables 7 and 8 answer the "*Can we use design parameters from a misspecified BIRD2 model (either intermediate or top level ignored) to plan a future BIRD3 design?*" question. Table 7 presents the misspecified BIRD2 model where the intermediate level was ignored. Power rates for a future BIRD3 design were calculated considering two cases. In the first case, one can use L2 parameters in the BIRD2 model for L3 parameters in the BIRD3 design (thus, L2 parameters in the BIRD3 design were all constrained to zero). In the second case, one can use L2 parameters in the BIRD2 model for L2 parameters in the BIRD3 design (thus, L3 parameters in the BIRD3 design were all constrained to zero).

When the intermediate level was ignored, considering case (i), calculated power rates slightly underestimated MC-based power rates for the misspecified BIRD2 model (see Table 7). They also underestimated MC-based power rates for the correctly specified BIRD3 model (see Table 1A in Appendix). However, in case (ii), calculated power rates were somewhat optimistic, substantially exceeding MC-based power rates of both BIRD2 and BIRD3 models (see Table 8 and Table 1A in Appendix). On the contrary, when the top-level was ignored, calculated power rates in case (i) were severely underestimated compared to both the BIRD2 model (see Table 8) and BIRD3 models (see Table 1A in Appendix), and in case (ii) they were unstable considering both models. The term "unstable" means we observed no trend regarding the magnitude or direction of the difference from MC-based power rates.

Table 7. 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 (use L2 parameters in the BIRD2 model for L3 parameters in the BIRD3 design)	0.33	0.24	0.67	0.44	0.38	0.22	0.59	0.40
(ii) `cosa` R Package (use L2 parameters in the BIRD2 model for L2 parameters in BIRD3 the design)	0.64	0.33	0.95	0.58	0.73	0.29	0.92	0.53

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 by L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained by 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*: The average number of L1 units per L2 unit was set to 100. *K*: Number of L3 units.

Table 8. 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 (use L2 parameters in the BIRD2 model for L3 parameters in the BIRD3 design)	0.20	0.20	0.23	0.19	0.23	0.18	0.20	0.18
(ii) `cosa` R Package (use L2 parameters in the BIRD2 model for L2 parameters in the BIRD3 design)	0.61	0.33	0.69	0.33	0.70	0.30	0.62	0.30

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 by L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained by 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*: The average number of L1 units per L2 unit was set to 20. *JK*: Number of L2 units. AD: Absolute difference. RD: Relative difference.

Discussion

This study investigated the consequences of ignoring either intermediate or top level on variance parameters and precision estimates in blocked three-level regression discontinuity (BIRD3) designs. There are various reasons to employ a misspecified model in this fashion (BIRD2 instead of BIRD3). The intermediate or top level information may be missing, the analysis may be too complex, or the researcher may be unaware of the consequences. Furthermore, BIRD2 models are common in practice; consequently, researchers may have no choice but to use parameters from BIRD2 models to plan for a BIRD3 design.

From an analysis perspective, when the intermediate level was ignored in the BIRD3 model, most of the variance in the ignored level shifted to the new bottom level, and a small portion of the variance shifted to the new top level. These results are in line with Moerbeek (2004), Van den Noortgate et al. (2005), and Opdenakker and Van Damme (2000). The shift in variance components causes a slight underestimation of power rates. It can be neglected if the variance of the intermediate level is small to moderate. This finding is in line with Zhu et al. (2011).

However, classroom-level variance can exceed school-level variance in practice (Goldstein, 2011; Muthen, 1991). One 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 indicates a substantial difference between the model that ignores and the model that acknowledges the intermediate level. In this case, it is advisable to acknowledge the intermediate level and pursue the analysis accordingly. Another way is to look at the L2 intra-class correlation coefficient. One could ignore the intermediate level if the intra-class correlation coefficient is small.

Since Type I errors did not change substantially when the intermediate level was ignored, one could use parameters from a misspecified BIRD2 model to plan for a BIRD3 design. The deterioration in the power rates will be negligible if L2 parameters in the misspecified BIRD2 model are used for L3 parameters in a future BIRD3 design. The top-level sample size could be oversampled by a few units to compensate for this. However, when L2 parameters of the misspecified BIRD2 model are used for L2 parameters in a future BIRD3 design, the test statistics will be underpowered during analysis. This approach should be avoided.

Ignoring the top level was more problematic, even with a small L3 variance. When the top level was ignored, the variance of the ignored level in the BIRD3 model shifted to the new top level, which is in line with

Moerbeek (2004) and Van den Noortgate et al. (2005). The shift in the variance reduces the power rate substantially; however, the increase in the sample size at the top level often compensates for this loss of power. Regardless, it should be avoided because Type I error rates were severely inflated.

Limitations

Results and their implications are limited to the simulated scenarios. Furthermore, ignoring a level may also mean omitting relevant variables at that level which introduces omitted variable bias. 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 do 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 as the 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). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. 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* (4th ed). 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*, *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*. Philadelphia: Consortium for Policy Research in Education. 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, Institute of Education Sciences, U.S. Department of Education. 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* (2nd ed). 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 A

Table 1A. Power rates for the correctly specified BIRD3 model

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
Power from `cosa` R Package	0.42	0.26	0.73	0.48	0.44	0.25	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 by L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained by L2 covariates. R_{T3}^2: Proportion of variance in the treatment effect explained by 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 was set to 20. *J*: Average number of L2 units per L3 units, which was set to 5. *K*: Number of L3 units.

Table 2A. Type I error rates for the correctly specified BIRD3 model

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
Type I Error from `cosa` R Package	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

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 by L1 covariates. R_{T2}^2: Proportion of variance in the treatment effect explained by L2 covariates. R_{T3}^2: Proportion of variance in the treatment effect explained by 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 was set to 20. *J*: Average number of L2 units per L3 units, which was set to 5. *K*: Number of L3 units.