Minimum Detectable Effect Size Computations for Cluster-Level Regression Discontinuity: Quadratic Functional Form and Beyond

Corresponding Author:
Metin Bulus
Department of Educational Measurement and Evaluation
College of Education, Adiyaman University, Adiyaman, Turkey
Email: bulusmetin@gmail.com
ORCID ID: orcid.org/0000-0003-4348-6322

October 2019

Abstract

Although Cattaneo, Titiunik, and Vazquez-Bare (2019) provides an ex-post data-driven framework for power computations in line with rdrobust Stata and R commands, which allows higher-order functional form of the score variable in non-parametric local polynomial estimation, ex-ante definitions for power computations are less clear and conventional definition of optimal design is not straightforward in cluster-level regression discontinuity (CRD) studies. This study extends power formulas proposed by Schochet (2008) assuming that the cluster-level score variable follows quadratic functional form. Results reveal that we need not be concerned with treatment by linear term interaction, and polynomial degree up to second order for symmetric truncation intervals. In comparison, every slight change in the functional form alters sample size requirements for asymmetric truncation intervals. Finally, an empirical framework beyond quadratic functional form is provided when the asymptotic variance of the treatment effect is untraceable. In this case, the CRD design effect is either computed from moments of the sample or approximate population moments via simulation. Formulas for quadratic functional form and the extended empirical framework are implemented in the cosa R package and companion Shiny web application.

Keywords: minimum detectable effect size, statistical power, clustered data, regression discontinuity design, polynomial functional form
Minimum Detectable Effect Size Computations for Cluster-Level Regression Discontinuity: Quadratic Functional Form and Beyond

Introduction

The birth of RDD dates back to Thistlewhite and Campbell (1960) within the context of education policy where they inquired into the effect of a merit-based scholarship program on student attitudes and career plans. Although practice of RDD was mostly abandoned in the subsequent several decades (Cook, 2008), re-adoption of the method in many fields have been catalyzed by several widely cited articles published in Volume 142 of the Journal of Econometrics - the historical narrative “Waiting for Life to Arrive” by Cook (2008), “A guide to practice” by Imbens and Lemieux (2008), Lee and Card’s (2008) article on specification bias, and McCrary’s (2008) article on density test for the score variable. About the same year, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences published a report on statistical power for RDDs in education evaluation (Schochet, 2008), and subsequently, the same ex-ante design issues were addressed in a journal article (Schochet, 2009).

Statistical power in multilevel RDDs is an understudied topic in the literature. Although studies and power analysis tools on multilevel randomized controlled trials (RCTs) abound, analogous advancements in multilevel RDDs have been lagging. In practice, we need ex-ante power analysis in RDDs to devise sufficiently powered studies to determine the effectiveness of an intervention targeting specific individuals, groups of individuals or organizations. In contrast to RCTs, selection procedure in RDDs is not random. Subjects, group of subjects or organizations are treated or withheld from the treatment based on criteria known as cutoff on a continuous scale, commonly referred to as running variable, forcing variable or score variable. For example, Dragoset et al. (2019) explored minimum detectable effect sizes (MDES) before data collection for a large-scale RDD to evaluate the effectiveness of School Improvement Grants (SIG) on student achievement. The SIG program provides funding to schools performing in the lowest quantile on academic proficiency tests. Schools that meet eligibility cutoff and receive funding are required to implement some of the rigorous interventions for improvement.

Some of the earliest known studies that acknowledged ex-ante design issues in RDDs were conducted by Goldberger (1972a, 1972b). Goldberger found that detecting treatment effect in simple individual-level regression discontinuity may require 2.75 as many subjects to reach the same level of precision as simple individual-level random assignment designs considering a normally distributed score variable. Schochet (2008, 2009) has extended this seminal work via investigating uniform, normal, truncated normal and bimodal score distributions within the context of multilevel modeling. What is notable and perhaps more applicable to education settings is that Schochet showed when bounds for the normal distribution is truncated around the cutoff, the sample size requirement in comparison to RCTs may increase by as much as 3 to 4 folds (as compared to 2.75). In practice these ratios, also known as regression discontinuity design effects (RDDE), could be much higher because score variable may not follow a normal distribution, may be skewed, or truncated with
(optimal) bandwidth selection procedures. Deke and Dragoet (2012) used empirical data from four large-scale RCTs to provide design effects. They considered Imbens and Kalyanaraman (2012) optimal bandwidth selection procedure and took into account specification error as suggested by Lee and Card (2008). It was shown that an RDD would have needed 9 to 17 times as many schools as RCTs to reach the same level precision. They note that, nonetheless, inflation in sample size requirement is largely driven by bandwidth selection as compared to specification error adjustment.

Not many large-scale evaluation studies utilize RDD, furthermore, not many of them report power analysis procedure explicitly nor do they refer to it. One possible reason could be that most popular statistical analysis procedures are data-driven, and power analysis procedures relying on data-driven techniques could be misleading as they are sample-based power estimates. Even though Schochet (2008, 2009) derived closed-form formulas for various multilevel RDDs, including two- and three-level designs with cluster-level discontinuity, derivations assume the linear functional form of the score variable. When the score variable is symmetric around the cutoff (default option in most of the non-parametric estimation routines), the majority of the time considering linear functional form alone may suffice for sample size planning.

Recently Cattaneo, Titiunik, and Vazquez-Bare (2019) elaborated on power computations and released Stata and R commands, taking mostly ex-post non-parametric local polynomial estimation (NLPE) perspective. Although the power computation routine allows higher-order functional forms, it is rarely needed because score distribution is truncated around the cutoff. The narrower the truncation interval gets, the more the relationship between the score variable and outcome becomes linear, obviating the need for higher-order functional forms. There are several shortcomings to this perspective. Ex-ante power definition is not clear, the conventional definition of optimal design is not straightforward, and whether the framework can be extended to the CRDs is not well understood (because the outcome and score variable are at different levels). A small simulation study using RDestimate (rdd v0.57, Dimmery, 2016), rdd_reg_np (rddtools v0.4.0, Stigler & Quast, 2015), and rdrobust (rdrobust v0.99.4, Calonico, Cattaneo, Farrell, & Titiunik, 2018) routines in R environment with default bandwidth selection procedures indicated power anomalies (mostly severely inflated Type I errors) when outcome and score variables are at different levels (results and R code available upon request). Furthermore, when generalizations to a larger population are of concern, we are more confident to make inferences by modeling outcome - score variable relationship across the full range of values, albeit it requires strict knowledge about the underlying functional form.

This does not mean including higher-order polynomials is always a good idea in conventional parametric mixed model estimation (PMME) where the full range of score variable is considered. Although including polynomial form ensures that any information on the selective assignment mechanism is partialed out from the estimate of future outcomes, higher-order polynomials may be unnecessary, they may even be harmful (Gelman & Imbens, 2017; Gelman & Zelizer, 2015). Gelman and Imbens (2017) put forward at least two related reasons as to why higher-order polynomials can be harmful. (i) Treatment effect estimate as a
function of the weighted average difference between treatment and control group depends on weights derived from the score variable and the cutoff. These weights can become extreme and distort discontinuity estimates beyond quadratic functional form. (ii) Polynomials beyond quadratic functional form produce erroneous confidence intervals that are prone to excluding zero, thus claiming a discontinuity effect in fact when there is none. Note that the second proposition is confirmed in the simulation study mentioned earlier. Strangely, Type I errors are inflated even with the linear functional form, to begin with, and they get worse with increased complexity in NLPE (results will be provided upon request). Finally, Deke, Wei, and Kautz (2017) brought our attention to small bias problems due to functional form misspecification in RDDs when researchers set out to devise studies that are capable of detecting a much smaller meaningful effect. Therefore, while the correct functional form is of crucial importance for devising studies sensitive to smaller effects, we may rarely need beyond quadratic functional form.

Taken together, advancements in the RDD methodology warrant reworking asymptotic variance of the treatment effect assuming the cluster level score variable follows quadratic functional form. In this way, more information about the score variable can be incorporated in the planning phase, and optimal design can be carried out where it is needed most. Fixed site effects are discussed through modifications to estimation strategy and degrees of freedom. Results are implemented in the cosa R library (Bulus & Dong, 2019), and the companion Shiny web application.

**Minimum Detectable Effect Size Computations**

The rationale for the minimum detectable effect (MDE) computation is to find the smallest effect that satisfies nominal Type I and Type II error rates have we known its standard error. In the context of regression discontinuity, Type I error (\(\alpha\)) is the rate of hypothetical samples that fail to detect treatment effect (\(\delta\)) when in fact there is an effect in the underlying population, whereas Type I error (\(\beta\)) rate is the rate of hypothetical samples that detect treatment effect when in fact there is none in the underlying population. Bloom (1995, p. 547) define MDE as “… the smallest effect that, if true, has X% chance of producing an impact estimate that is statistically significant at Y level.” where X is the power rate (1 – \(\beta\)) and Y is the \(\alpha\) level. When standardized in Cohen’s \(d\), this smallest effect is referred to as the minimum detectable effect size (MDES, Bloom, 2006), which is comparable across samples and time.

MDES can be computed given Type I and Type II error rates (and degrees of freedom for small samples) has there been a prior for standard error, which is more interpretable and intuitive in comparison to statistical power. MDES for a two-tailed test can be computed as

\[
MDES(\delta) = \left( t_{\alpha/2, v} + t_{1-\beta, v} \right) \sqrt{\frac{\sigma_\delta^2}{\sigma_Y^2}}
\]

where \(t\) is the quantile function for student’s t-distribution, \(v\) is degrees of freedom, \(\alpha/2\) is Type I error rate for two-tailed hypothesis testing (\(\alpha\) for one-tailed) and \(\beta\) is Type II error rate, \(\sigma_\delta^2\) and \(\sigma_Y^2\) are variances of the treatment effect and the outcome. The term with the
The square root is the standardized standard error of the [standardized] treatment effect and can be stated in terms of known design parameters such as expected sample size, intra-class correlation coefficients, and R-squared values. The goal of this study is to derive closed-form formulas and provide an extended empirical framework for standardized standard errors so that researchers can set priors given design parameters.

The rest of this article is organized in several sections as follows. In the first section, I elaborate on statistical models, derive the asymptotic variance of the cluster-level discontinuity effect, and define parameters within the formula. I also elaborate on the analytics of moment-based approach to finding population-based correlations between treatment condition, score variable, and quadratic form of the score variable. In the second section, I closely inspect and elaborate on the properties of the derived formulas. In the third section, I consider some of the minor albeit important design issues commonly raised by researchers and practitioners. In the fourth section, I briefly introduce R functions and implement them in an illustrative example that is based on a recent evaluation report in the fifth section. Finally, I conclude the study, discuss the advantages and disadvantages of NLPE and PMME approaches to power computations, and offer some future directions for this line of research.

Statistical Models and Derivations

In this section, I present statistical models for two-level CRDs (hereafter, CRD2), derive closed-form formula for the asymptotic variance of the treatment effect, and standardize the formula in terms of intra-class correlation coefficient and R-squared values. I use the terms "subjects" to refer to level 1, and "clusters" to refer to level 2, but also sometimes use them interchangeably. Statistical models and derivations for three- and four-level CRDs are available in the Supplement.

Unconditional Model for CRD2 Design

The following unconditional model is used to obtain variance parameters $\sigma^2$ and $\tau^2$ as defined below, which will be used to obtain intra-class correlation coefficient and R-squared values along with parameters obtained from the full model.

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

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

where $r_{ij} \sim N(0, \sigma^2)$, $\mu_{0j} \sim N(0, \tau^2)$, and $\sigma^2$ and $\tau^2$ are variances for level 1 and level 2 residuals sum of which virtually adds up to the outcome variance. The null model states that the outcome for subject $i$ in cluster $j$ ($Y_{ij}$) is the sum of the grand mean ($\gamma_{00}$), a random effect associated with the cluster ($\mu_{0j}$), and a random effect associated with the subject ($r_{ij}$). In this way variation in the outcome is partitioned into within and between clusters. This will be useful to standardize conditional residual variances obtained from the full model below.

Full Model for CRD2 Design
The full model partitions conditional residual variance into within and between clusters. Conditional residual variances depend on variables introduced at level 1 and level 2. Cluster level treatment variable ($T_j$) is derived from the cluster level score variable ($Z_j$) based on the cutoff ($Z_0$). Score variable is centered around the cutoff and modeled up to second order along with the treatment variable, level 1 and level 2 covariates ($X_{ij}$ and $W_j$). The full model states that outcome for subject $i$ in cluster $j$ ($Y_{ij}$) is sum of the intercept ($\gamma_{00}$), additive terms multiplied with their respective regression coefficients ($\gamma_{01}$ for treatment effect, $\gamma_{02}$ and $\gamma_{03}$ for polynomial functional form of the score variable, $\gamma_{10k}$ and $\gamma_{04}$ for level 1 and level 2 covariates), a random effect associated with the cluster ($\mu_{0j}$), and a random effect associated with the subject ($r_{ij}$).

The full model is used to obtain variance parameters $\sigma_{1|x}^2$ and $\tau_{1|x,Z,W}^2$ as defined below, which are used to calculate R-squared values along with the parameters from the unconditional model.

**Level 1:**

$$ Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} $$

**Level 2:**

$$ \beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}(Z_j - Z_0) + \gamma_{03}(Z_j - Z_0)^2 + \gamma_{04}W_j + \mu_{0j} $$

$$ \beta_{1j} = \gamma_{10k} $$

where $r_{ij} \sim N(0, \sigma_{1|x}^2)$, $\mu_{0j} \sim N(0, \tau_{1|x,Z,W}^2)$, intra-class correlation coefficient is defined as $\rho = \tau^2 / (\tau^2 + \sigma^2)$ and represents proportion of variance in the outcome between level 2 units, $\sigma_{1|x}^2$ and $\tau_{1|x,Z,W}^2$ are level 1 and level 2 residual variances conditional on predictors at their respective level, R-squared value for level 1 is defined as $R_1^2 = 1 - \sigma_{1|x}^2 / \sigma^2$ and is proportion of level 1 variance in the outcome explained by level 1 predictors, R-squared value for level 2 is defined as $R_2^2 = 1 - \tau_{1|x,Z,W}^2 / \tau^2$ and is proportion of level 2 variance in the outcome explained by level 2 predictors.

*Derivation of Asymptotic Variance for Discontinuity Effect in CRD2 Design*

The generalized least squares estimate of the covariance matrix for fixed effect coefficients can be stated as

$$ \text{cov}(\gamma) = \left( \sum_{j=1}^{J} X_j^T V_j^{-1} X_j \right)^{-1} $$

where $X_j$ is $n \times 4$ design matrix for cluster $j$ including intercept, $T_j$, $Z_j$, $Z_j^2$ as

$$ X_j = \begin{bmatrix} 1_{n_P} & (1 - P)1_{n_P} & (Z - Z_0)n_P & (Z - Z_0)^2n_P \\ 1_{n(1 - P)} & -P1_{n_P} & (Z - Z_0)n_{(1 - P)} & (Z - Z_0)^2n_{(1 - P)} \end{bmatrix} $$

and $n$ is the number of level 1 units per cluster on average, $P$ is the proportion of level 1 units below (or above) cutoff. This design matrix is same for each cluster because treatment and score variable is at the cluster level. $V = ZGZ^T + R$, $R$ is $(nJK) \times (nJK)$ diagonal level 1
residual variance matrix, \( G \) is \( J \times J \) diagonal level 2 residual variance matrix, \( Z \) is \((nJ) \times (J \times 1)\) matrix of random effects (note that \( x1 \) is due to having only intercept random across level 2 units). In more complicated models \( Z \) can be obtained via \( Z = J^T \otimes X^T \) where \( J \) is the indicator (sparse) matrix of group membership for each observation and \( X \) is the design matrix for random effects. The resultant \( V \) is \((nJ) \times (nJ)\) block diagonal sparse matrix, in which within each block diagonal structure represents the total residual variance and off diagonals are level 2 residual variance. For example, assume \( n = 2 \) and \( J = 2 \) then diagonal elements of \( V \) matrix for unconditional model takes the form

\[
V_j = \begin{bmatrix}
\sigma^2 + \tau^2 & \tau^2 \\
\tau^2 & \sigma^2 + \tau^2
\end{bmatrix}
\]

Assuming constant variance and compound symmetry structure, only intercept is random across level 2 so \( Z_j \) is a vector of 1s with a length of \( n \), so \( V_j = 1_n^T \sigma^2_{|Z} Z_j^2 W j^T + \sigma^2_{|X} I_n \), where \( 1_n \) in bold indicate vector of 1s with length \( n \), and \( I_n \) is \( n \times n \) identity matrix. Then

\[
cov(\gamma) = \frac{(n\tau^2_{|Z} Z_j^2 W + \sigma^2_{|X}) (X^T X)^{-1}}{J}
\]

The cell associated with the treatment effect converge in probability to

\[
cov(\gamma)_{(2,2)} \rightarrow \frac{n\tau^2_{|Z} Z_j^2 W + \sigma^2_{|X}}{Jn \rho (1 - \rho)} \left( \frac{1 - \rho_{zz}^2}{1 - \rho_{zz}^2 - \rho_{zz}^2 - \rho_{zz}^2 + 2\rho_{xz} \rho_{xz} \rho_{zz}^2} \right) \quad \#(1)
\]

Equation 1 can be stated in standardized form since we know \( \rho = \tau^2 / (\tau^2 + \sigma^2) \), \( R_1^2 = 1 - \frac{\sigma^2_{|X}}{\sigma^2} \) and \( R_2^2 = 1 - \frac{\tau^2_{|Z} W}{\tau^2} \):

\[
var(\gamma)_{01} = \frac{n \rho (1 - R_2^2) + (1 - \rho)(1 - R_1^2)}{Jn \rho (1 - \rho)} \left( \frac{1 - \rho_{zz}^2}{1 - \rho_{xz}^2 - \rho_{xz}^2 - \rho_{zz}^2 + 2\rho_{xz} \rho_{xz} \rho_{zz}^2} \right) \quad \#(2)
\]

In Equation 2, the treatment effect is reparametrized in terms of Cohen’s \( d \) as \( \gamma_{01} = \gamma_{01} / \sqrt{\sigma^2 + \tau^2} \). It can be seen that inflation in variance due to selective treatment assignment in contrast to CRTs is a function of correlations between treatment indicator, score variable, and its quadratic form. Hereafter I will refer to the multiplier on the right side of the formula as regression discontinuity design effect (RDDE) which is the variance inflation factor in comparison to CRTs. Throughout the text, RDDE is interchangeably used as the ratio of sample size in the CRD to corresponding CRTs to reach the same level of precision. Perhaps it might worth to mention that for multisite RDDs this inflation factor only inflates variance portion associated with discontinuity level and below, so this statement should be generalized with some caution. In the next section, I detail analytics of computing correlations based on population moments when we do not have an empirical score variable at our disposal. Alternatively, they can be simulated.

**Correlations Between \( T \), \( Z \) and \( Z^2 \) Triad**
It was established that RDDE is a function of correlations between $T$, $Z$ and $Z^2$ triad when quadratic functional form is considered. When a score variable is available these correlations can be computed empirically, but this would produce sample-based MDES values. To obtain population-based MDES values we should either derive analytic forms of correlations based on population moments, or conduct simulations to approximate population moments. Although Schochet (2008) consider uniform, normal, truncated normal and bimodal distributions to derive analytic forms of the correlation between treatment condition and linear score variable as a function of $p$, I only consider the uniform and truncated normal distributions because we can define lower and upper bounds for bandwidth considerations. Furthermore, widening lower ($k_1$) and upper ($k_2$) bounds for truncated normal distribution reasonably approximate normal distribution to produce accurate correlation coefficients (e.g. $k_1 = -20$ and $k_2 = 20$). As for more complex distributions, especially because the score variable is known in advance, correlations can be computed empirically based on sample moments.

Regardless of the form of the distribution, correlation between $T$, $Z$ and $Z^2$ triad can be computed using a moment-based approach. Let $M_k()$ be moment generating function for order $k$. Because definitions for variances follow

$$
s_T^2 = P(1-P)$$
$$
s_Z^2 = M_2(Z) - M_1^2(Z)$$
$$
s_{Z^2}^2 = M_4(Z) - M_2^2(Z)
$$

then correlations

$$
\rho_{TZ} = \frac{\sigma_{TZ}}{\sigma_T\sigma_Z} = \frac{P[M_1(Z|Z \geq Z_0) - M_1(Z)]}{\sqrt{P(1-P)[M_2(Z) - M_1^2(Z)]}} \tag{3}
$$

$$
\rho_{TZ^2} = \frac{\sigma_{TZ^2}}{\sigma_T\sigma_{Z^2}} = \frac{P[M_2(Z|Z \geq Z_0) - M_2(Z)]}{\sqrt{P(1-P)[M_4(Z) - M_2^2(Z)]}} \tag{4}
$$

$$
\rho_{Z^2} = \frac{\sigma_{TZ}}{\sigma_Z^2} = \frac{M_3(Z) - M_1(Z)M_2(Z)}{\sqrt{[M_2(Z) - M_1^2(Z)][M_4(Z) - M_2^2(Z)]}} \tag{5}
$$

where for uniform distribution analytic form of the moment generating function is

$$
M_k(Z) = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)} \tag{6}
$$

where $a$ and $b$ are lower and upper bounds. For truncated normal distribution moment generating function follows a recursive form as (Burkardt, 2014; Horrace, 2015)
\[ M_k(Z) = (k - 1)\sigma_Z^2 M_{k-2}(Z) + \mu_Z M_{k-1}(Z) - \sigma_Z \frac{b^{k-1} \phi \left( \frac{b - \mu_Z}{\sigma_Z} \right) - A^{k-1} \phi \left( \frac{a - \mu_Z}{\sigma_Z} \right)}{\Phi \left( \frac{b - \mu_Z}{\sigma_Z} \right) - \Phi \left( \frac{a - \mu_Z}{\sigma_Z} \right)} \]  

where \( M_0(Z) = 1, A = (a - \mu_Z)/\sigma_Z, B = (a - \mu_Z)/\sigma_Z, a \) and \( b \) are lower and upper bounds for \( Z \sim N(\mu_Z, \sigma_Z^2) \), \( \phi \) is probability density function and \( \Phi \) is cumulative density function for standard normal distribution. Note that conditional moment in the numerator of Equations 3, 4 and 5 can be obtained via altering lower or upper bounds in the moment generating function (Equation 6 or 7) depending on whether the treatment group is above or below the cutoff score respectively. For example, subjects scoring above a cutoff may be assigned to treatment group in a scholarship program. On the contrary, subjects scoring below a cutoff may be assigned to a treatment condition in a remedial program. Although the direction of treatment status can alter correlations (from positive sign to negative or vice versa) it does not influence MDES computations.

We do not demonstrate directionality of treatment assignment in notations but provide this option in R functions, for which correlations are computed either based on moments analytically, using empirical data, or via simulation. Moments of the truncated normal distribution are derived from (Burkardt, 2014) for analytic computations. Simulations for the truncated normal distribution is based on the accept-reject algorithm described in Robert (1995) and implemented in \texttt{msm} R package (Jackson, 2011). The default is based on analytic solutions up to second order, however, if requested, results from the simulations are based on a sample size of 1000 averaged over 1000 repetitions. Users should be aware that only RDDE is simulated in the variance function, not the entire RDD model because the outcome variable is assumed to be unavailable during the planning phase, which is consistent with the evaluation theory.

**Distribution Type, Skewness, and Truncation Interval**

A more detailed inspection reveals interesting properties of derived formulas. Table 1 and 2 inspect the influence of distribution type, functional form and asymmetric truncation intervals on the number of clusters in comparison to CRTs in light of correlations between \( T, Z \), and \( Z^2 \) triad. Figure 1 to 3 adds to this information in that they focus on RDDE, consider the proportion of clusters in the treatment group on a continuous scale, and indicate the intersection points of RDDE lines for linear and quadratic functional forms.

Table 1 and 2 are interpreted under two groups of scenarios; in the first (shaded rows), distributions are symmetric around zero mean. We will see that in such cases, some of the correlations between \( T, Z \) and \( Z^2 \) triad are non-existent. In the next group of scenarios truncated normal distribution is asymmetric, and uniform distribution has a non-zero mean.

Table 1 and 2 are revealing in the sense that for symmetric truncated normal or uniform distributions some of the correlations between \( T, Z \), and \( Z^2 \) triad are zero. Regardless of the cutoff or \( p \), the correlation between \( Z \) and \( Z^2 \) is zero. Besides, when \( p = .50 \), the correlation between \( T \) and \( Z^2 \) is also zero (see Table 1 and 2, and Figure 1 and 2). Although
correlations remain unchanged with varying truncation intervals in a centered uniform distribution, they differ in symmetric truncated normal distribution. It is also apparent in Tables 1 and 2 that complementary values of $p$ (.10 vs .90, .40 vs .60, etc.) produce the same number of clusters in the first group of scenarios.

Asymmetric truncated normal or uniform distributions have shifted peak points for the correlation between $T$ and $Z$ in comparison and shifted inflection point for the correlation between $T$ and $Z^2$ as a function of $p$ (see Figure 3). Correlation between $T$ and $Z$ does not depend on interval width in a uniform distribution, however, the correlation between $T$ and $Z^2$, and $Z$ and $Z^2$ are affected by the interval width (Table 1 and 2). Correlation between $T$ and $Z^2$ is zero below $p = .50$ for right-skewed distribution and above $p = .50$ for left-skewed distribution (Figure 3). Correlation between $Z$ and $Z^2$ is non-zero positive for right-skewed and non-zero negative for left-skewed distribution (Figure 3). Contrary to the first group of scenarios, complementary values of $p$ (.10 vs .90, .40 vs .60, etc.) produce different numbers of clusters for the same truncation points.

[Tables 1 and 2 about here]

For a CRD with a normally distributed score variable, we may need up to 2.75 times as many clusters as the corresponding CRT when the linear functional form is considered, and up to 2.97 times as many clusters when the quadratic functional form is considered. The peek ratio occurs at $p = .50$ for the linear functional form and $p \approx .21$ or $p \approx .79$ for the quadratic functional form. The peek ratio for linear functional form and saddle point for quadratic function intersects at $p = .50$. At this point adding the quadratic term is not associated with a higher sample size requirement. These ratios are even higher when a distribution is truncated and/or asymmetric. For example, when the score variable is truncated between one standard deviation below and above mean, we may need up to 3.65 times as many clusters as the corresponding CRT when the linear functional form is considered and 4.58 times as many clusters when the quadratic functional form is considered. The peek ratio occurs at $p = .50$ for the linear functional form and $p \approx .25$ or $p \approx .75$ for the quadratic functional form. When normal distribution is asymmetric the peak ratio for the linear functional form and saddle point for the quadratic function intersects somewhere other than $p = .50$.

For a CRD with a uniformly distributed score variable, we may need up to 4.00 times as many clusters as the corresponding CRT when the linear functional form is considered, and up to 5.00 times as many clusters when the quadratic functional form is considered. The peek ratio occurs at $p = .50$ for the linear functional form and $p \approx .28$ or $p \approx .72$ for the quadratic functional form. The narrower the interval for a truncated normal distribution the more it resembles a uniform distribution, thus, upper limits for a truncated normal distribution with unknown truncation points approximates upper limits for a uniform distribution where we may need up to 4.00 times as many clusters as the corresponding CRT when the linear functional form is considered and up to 5.00 times as many clusters when the quadratic functional form is considered (compare Figures 1 and 2).

[Figures 1, 2 and 3 about here]
As seen in Figures 1 to 3, there is a point where RDDE does not change depending on functional form (up to quadratic term without interactions). This point is at $p = .50$ for the symmetric case, which is the intersection of peak point for the linear and saddle point for the quadratic line in Figures 1 and 2 but diverge from $p = .50$ otherwise. In the latter case, the intersection of the linear and quadratic lines in Figure 3 is not exactly the peak point and saddle point. The case of intersection point worth paying attention because it may partially explain why including higher-order terms (which are not correlated with the treatment or each other) might produce noisy estimates. Thus, if a quadratic term is not correlated with the outcome conditional on the linear term, and $p$ happens to be near the intersection point, the quadratic term has no added value to noise reduction.

**Design Considerations**

In this section, I elaborate on some of the RDD design issues commonly raised by researchers. Throughout, I embrace the idea “plan for the worst, hope for the best”. Plan for the worst means considering the worst-case scenario when we have multiple options for design parameters, hoping for the best means hoping that ex-post empirical MDES values are equal or lower than what we had expected. Understandably, this may not be always possible due to resource constraints, thus, some of the headings below become relevant.

**Attrition, Treatment Compliance, and Crossovers**

Recently, Rickles, Zeiser, and West (2018) raised the issue of attrition, provided some benchmarks and guidelines in this regard. In this study, similar to CRTs as noted by Rickles et al. (2018), level 1 attrition may not be as important as level 2 attrition in CRD2. The rule of thumb is, we are less likely to afford attrition for levels closer to where discontinuity resides. For example, within the education context, bets should be on considering attrition at the school level rather than students in the planning phase. This notion does not carry over to multisite RDDs, in which student attrition would be a real concern. Rickles et al. (2018) reminds us that attrition is just another design parameter and that it should be reported ex-post and considered ex-ante. I do not explicitly incorporate attrition in MDES computations. As long as attrition rates are established based on prior research or empirical data, adjustments are straight forward in CRDs.

Unlike attrition, treatment compliance and control crossovers are accounted for with the estimation procedure, known as fuzzy estimation in RDDs. In fuzzy RDD, estimation involves complier average or local average treatment effect (LATE), accounting for treatment group participant rate ($r_{TP}$) and control group crossovers ($r_{CC}$) (Imbens & Angrist, 1994; Hanh, Todd, & Klaauw, 2001; Schochet, 2008). To adjust for $r_{TP}$ and $r_{CC}$ in power computations Schochet (2008) has established that

$$\text{var}(\gamma_{01}(\text{LATE})) = \text{var}\left(\frac{\gamma_{01}}{r_{TP} - r_{CC}}\right) = \text{var}(\gamma_{01}) \left(\frac{r_{TP} - r_{CC}}{r_{TP} - r_{CC}}\right)^2 \#(8)$$

using Tylor series expansion and assuming $r_{TP} - r_{CC}$ as a fixed design parameter (thus only the first term of the Tylor series applies). In reality, $r_{TP} - r_{CC}$ in Equation 8 is estimated based
on regressing observed treatment status on $T$ and $Z$, which has its uncertainty that ought to be taken into consideration (Schochet, 2008). We adopt the earlier case and also assume $r_{TP} - r_{CC}$ is a fixed design parameter similar to standardized variance parameters used in the approximate variance formula. It might be hardly the case that a school in the control group would switch to receive treatment, though not entirely impossible, as we will see in an illustrative example.

**Bandwidth Considerations and External Validity**

In RDD non-parametric local polynomial estimation (NLPE) score variable is truncated with bandwidth selection procedures to obtain an unbiased estimate of the treatment effect (Calonico, Cattaneo, & Farrell, 2018; Imbens & Kalyanaraman, 2012) despite inflated standard errors, reduced power, and lower external validity. NLPE makes the model robust to functional form specification in addition to other advantages. External validity is a real concern in NLPE for which various strategies have been offered for its evaluation and improvement in the literature (see, Andrews & Oster, 2018; Bertanha & Imbens, 2019; Cattaneo, Keele, Titinuik, & Vazquez-Bare, 2019; Cerulli, Dong, Lewbel, & Poulson, 2017; Wing & Bello-Gomez, 2018). In its simplest form, it is possible to improve external validity via considering the full range of the score variable assuming the functional form is correctly specified. When functional form is correctly specified, this form of estimation with PMME has lower standard errors, superior power rates, and produce estimates as correct as NLPE. PMME can tolerate slight deviations from normality.

Optimal bandwidth considerations apply to data-driven approaches when the outcome variable is available in advance. If this is the case, users can use a software of the choice such as rdpower and rdrobust Stata and R packages (Calonico, Cattaneo, Farrell, & Titinuik, 2018; Cattaneo, Titinuik, & Vazquez-Bare, 2019). rdpower is a promising tool for power analysis of RDDs, and it seems it is appropriate when the score variable and outcome are at the same level. Influence of cluster-level discontinuity based on cluster-level score variable when the outcome is available at lower levels has not been studied in-depth, and very little is mentioned in the manual.

**Explanatory Power of Covariates**

One kind of parameter that needs to be known before an evaluation study is R-squared values representing a standardized measure for the explanatory power of covariates. Although it is best to assume there are no covariates to comply with “plan for the worst, hope for the best” notion, with larger sample size and prohibitive costs researchers steered towards the optimal design of multilevel RCTs and RDDs (Bulus & Dong, 2019; Hedges & Borenstein, 2014; Konstantopoulos, 2009, 2011, 2013; Liu, 2003; Moerbeek, & Safarkhani, 2018; Raudenbush, 1997; Raudenbush & Liu, 2000; Rhoads & Dye, 2016). It has been shown that one effective way to reduce the required sample size it to acknowledge the fact that at least minimal covariates will be collected and accounted for in the future analytic models (Bloom, Richburg-Hayes, Black, 2005; Bulus & Sahin, 2019). There is an emerging body of literature undertaking this task for various outcome measures spanning to a wide range of geographical area for different education systems (e.g., Brunner et al., 2017; Dong et al., 2016; Hedges &
Hedberg, 2013; Juras, 2016; Spybrook, Westine, Taylor, 2016; Westine, Spybrook, Tylor, 2014). Although these studies mostly focus on CRTs within the context of multilevel modeling, results can be carried over to CRDs, and to NLPE (see Calonico, Cattaneo, Farrel, & Titiunik, 2019).

Consideration of the explanatory power of covariates is important in CRD studies as it would require many more clusters and participants compared to corresponding CRTs. While studies reporting standardized variance parameters assume minimal covariate information, mostly pretest and demographic variables, there might be cases where stakes should be placed on collecting more information. One reason is that it would improve the power rate, counteracting the deterioration of power in the face of attrition and crossover; another is that budgetary constraints may prevent from sampling more clusters and subjects.

Derivatives of the variance function concerning level 1 and level 2 R-squared values provide the direction and magnitude of change in the variance induced by changes in R-squared values. We elaborate on this issue for CRD2 and CRD3 but results carry over to CRD4.

Two-level CRD

Earlier it was established that RDDE is a function of correlations between $T$, $Z$, and $Z^2$ triad. To save space I will write RDDE in parenthesis in variance functions. The standardized variance function for CRD2 takes the form

$$\text{var}(\gamma_{01}^*) = \frac{(\text{RDDE})\rho_2(1 - R_2^2)}{JP(1 - P)} + \frac{(\text{RDDE})(1 - \rho_2)(1 - R_1^2)}{nJP(1 - P)}$$

and derivatives with respect to $R_1^2$ and $R_2^2$ are

$$\frac{\partial \text{var}(\gamma_{01}^*)}{\partial R_1^2} = \frac{-(\text{RDDE})(1 - \rho_2)}{nJP(1 - P)} \quad \#(9)$$

$$\frac{\partial \text{var}(\gamma_{01}^*)}{\partial R_2^2} = \frac{-(\text{RDDE})\rho_2}{JP(1 - P)} \quad \#(10)$$

Both derivatives in Equations 9 and 10 have a negative sign, indicating that an increase in R-squared values will induce a decrease in the variance. Derivatives are similar to what we would one obtain for two-level CRTs (Bulus & Sahin, 2019), however, they differ by RDDE multiplier. Interestingly, this means including covariates in CRD2 studies reduces variance RDDE times more which emphasizes the importance of including covariates in CRDs.

Assume a typical scenario in education where the treatment condition is determined based on the lowest quantile of cluster-level score variable, which follows a standard normal distribution and takes quadratic functional form, further assume other design parameters are $\rho_2 = .20, R_1 = .50, R_2 = .50, n = 20, \text{and } J = 150$. I will increase the R-squared values by
.10 one at a time and observe the change in the variance and power rates. For example, re-specifying $R_1^2 = .60$ marginally decreases variance from 0.0149 to 0.0144 which translates into a marginal change in power rate from .5311 to .5449, whereas re-specifying $R_2^2 = .60$ decrease variance to 0.0124 which increase power rate to .6073. Thus, attempts to improve the explanatory power of covariates at the discontinuity level are typically rewarding. This generalization applies to typical scenarios in education, exceptions exist, albeit not typical, for example, if $n = 1$ and $\rho < .50$ increasing $R_2^2$ decrease variance or increase power rate less than increasing $R_1^2$ by the same amount.

Three-level CRD

The standardized variance function for CRD3 takes the form

$$\text{var}(\xi_{100}) = \frac{(RDDE)\rho_3(1 - R_3^2)}{KP(1 - P)} + \frac{(RDDE)\rho_2(1 - R_2^2)}{JKP(1 - P)} + \frac{(RDDE)(1 - \rho_2 - \rho_3)(1 - R_1^2)}{nJKP(1 - P)}$$

and derivatives with respect to $R_1^2$, $R_2^2$, and $R_3^2$ are

$$\frac{\partial \text{var}(\xi_{100})}{\partial R_1^2} = -\frac{(RDDE)(1 - \rho_2 - \rho_3)}{nJKP(1 - P)} \tag{11}$$

$$\frac{\partial \text{var}(\xi_{100})}{\partial R_2^2} = -\frac{(RDDE)\rho_2}{JKP(1 - P)} \tag{12}$$

$$\frac{\partial \text{var}(\xi_{100})}{\partial R_3^2} = -\frac{(RDDE)\rho_3}{KP(1 - P)} \tag{13}$$

Similar to CRD2, all three derivatives in Equations 11, 12, and 13 have negative signs which indicate increasing R-squared values decrease variance. It seems as we get closer to the discontinuity level, typically a change in R-squared value induce a larger change in variance, as seen from denominators. Again, derivatives are similar to what we would obtain for three-level CRTs (Bulus & Sahin, 2019), however, they differ by the RDDE multiplier. This also means including covariates in CRD3 studies reduces variance RDDE times more in comparison to corresponding three-level CRT.

Similar to CRD2, I will increase the R-squared values by .10 one at a time and observe the change in the variance and power rates. For example, assume the same properties for score variable and cutoff as in CRD2, and further assume design parameters $\rho_2 = .20$, $\rho_3 = .10$, $R_1^2 = .50$, $R_2^2 = .50$, $R_3^2 = .50$, $n = 20$, $J = 3$ and $K = 150$. Re-specifying $R_1^2 = .60$ decreases variance marginally from 0.0110 to 0.0109 which translates into a marginal increase in power rate from .6569 to .6627. In comparison, re-specifying $R_2^2 = .60$ decreases variance to 0.0102 and power rate to .6907, whereas re-specifying $R_3^2 = .60$ decreases variance to 0.0098 and power rate to .7085. Again, exceptions exist, albeit atypical.

**Beyond Quadratic Functional Form: Higher Orders, Interactions**
As we increase the complexity of the model, it becomes more and more difficult to track down the sampling variance of the treatment effect. In other words, \((X^TX)^{-1}\) is untraceable. To overcome this challenge we can derive RDDE via simulating only \((X^TX)^{-1}\) part of the formula multiple times (rather than simulating the full model) based on the distribution of the score variable, its functional form and whether it interacts with the treatment. Currently only truncated normal and uniform distributions are allowed for which bounds can be modified by the user. When there is an empirical score variable, often there is, \((X^TX)^{-1}\) can be computed empirically.

[Table 4 about here]

Considering linear form, its interaction with the treatment indicator or quadratic functional form does not matter for symmetric truncation points (they all intersect at \(p = .50\)). However, RDDE jumps when the cubic form is added (see Table 4), or when interactions with quadratic and cubic forms are considered. This can be interpreted in two ways, (i) we need not be concerned about the interaction with the linear term, or the quadratic term in the model when truncation interval is symmetric around the cutoff (assuming the true model is any of the three distinct specifications), (ii) failure to model cubic form could be a grave mistake had it been the true functional form within the interval. This interpretation does not apply to asymmetric truncation intervals where an increase in complexity in any form increases RDDE. However, there is no foolproof method to identify the underlying functional form (Gelman & Imbens, 2018). Thus, functional form within asymmetric intervals should be chosen with caution.

*(Conditional) Optimal Design*

I refrain from listing (conditional) optimal design functions (ODFs) here, as they are same as those reported in the literature. For the CRD2 design, ODFs are same as functions derived in Rhoads and Dye (2016). As Rhoads and Dye (2016) noted, ODFs in CRD2 design is not influenced by the functional form or their interactions with the treatment indicator. In fact, ODFs are same as two-level CRTs. For CRD3 and CRD4 designs, ODFs are same as three- and four-level hierarchical random assignment designs in Hedges and Borenstein (2014). In other words, functional form in CRDs bears no weight when total cost is fixed. Thus, ODFs for CRTs or CRDs are the same.

Conventional analytic formulas for optimal design of CRTs and CRDs are rarely used in practice, whereas finding cost-effective allocation with precision constraints may be more practical (Bulus & Dong, 2019). In this case, when the total cost is minimized with constraints placed on MDES or power rate, results between CRTs and CRDs no longer match, and slight modifications to functional form produce different results. Formulas for quadratic functional form and empirical extension is incorporated into the Bound Constrained Optimal Sample Size Allocation (BCOSSA) framework proposed by Bulus and Dong (2019a) and implemented in cosa R package (Bulus & Dong, 2019b).

**Software Implementation**
Functions to compute MDES for CRD2 and CRD3 designs are implemented in the cosa R package or companion Shiny web application. MDES functions for CRD2 and CRD3 designs and their arguments are presented below (can be extended to CRD4 design – omitted to save space).

```r
mdes.crd2(score = NULL, dists = "normal", k1 = -6, k2 = 6, 
order = 2, interaction = FALSE, treat.lower = TRUE, cutoff = 0, 
power = .80, alpha = .05, two.tailed = TRUE, 
df = n2 - g2 - order * (1 + interaction) - 2, 
rho2, r21 = 0, r22 = 0, g2 = 0, 
rate.tp = 1, rate.cc = 0, p = NULL, n1, n2)

mdes.crd3(score = NULL, dists = "normal", k1 = -6, k2 = 6, 
order = 2, interaction = FALSE, treat.lower = TRUE, cutoff = 0, 
power = .80, alpha = .05, two.tailed = TRUE, 
df = n3 - g3 - order * (1 + interaction) - 2, 
rho2, rho3, r21 = 0, r22 = 0, r23 = 0, 
g3 = 0, rate.tp = 1, rate.cc = 0, p = 'NULL, n1, n2, n3)
```

Arguments in the first two lines are pertinent to the score variable. `score` argument can be an empirical score variable, or an object inheriting class “score” (produced by `inspect.score()` function which has first two lines as its argument). Submitting an object of class “score” obviates the need for specifying other arguments in the first two lines. They only become active when no score variable is provided (default). Assuming no score variable is provided, `dists` argument refers to the score distribution, currently accepts "normal" or "uniform". `k1` and `k2` are lower and upper limits for score distribution. `order` argument is the polynomial order for the score variable and `interaction` is a logical argument, if true, meaning score variable interacts with the treatment indicator. `treat.lower` argument is logical, if true, clusters below cutoff are treated, and `cutoff` is threshold or cutoff. `treat.lower` does not change results, but is kept for future compatibility.

`power` is power rate, `alpha` is Type I error rate, and `two.tailed` is logical, if true, hypothesis test is based on two-tailed test. `df` is degrees of freedom and is derived from other arguments, but can be modified by the user to consider fixed block effects. `rho2` and `rho3` are intra-class correlation coefficients for level 2 and level 3, `r21`, `r22` and `r23` are R-squared values associated with level 1, level 2, and level 3 covariates. `g2` and `g3` are number of covariates included in the model at level 2 and level 3 other than treatment indicator, score variable and their interaction. `rate.tp` is treatment group participation rate, and `rate.cc` and is control group cross-over rate. `p` is proportion of clusters below or above cutoff which is just an alternative to `cutoff` argument. `n1` is average number of subjects per cluster, `n2` is number of clusters (or average number of level 2 sub-clusters per level 3 clusters in CRD3), and `n3` is number clusters. Among all the arguments, for functions to successfully run with defaults, at least `rho2`, `n1`, and `n2` should be specified in CRD2 designs, and at least `rho2`, `rho3`, `n1`, `n2`, and `n3` should be specified in CRD3 designs.

**Illustrative Example**

In this section, I will briefly describe a recent large-scale evaluation study that used RDD with a cluster-level score variable and continue as if I am planning another study of similar kind. Thus, this illustration should not be taken as a critic for MDES computations in
the evaluation study - from which I merely draw design parameters - but rather as a future study for which I have access to all resources - monetary and otherwise.

Dragoset et al. (2019) use RDD to evaluate the effectiveness of School Improvement Grants (SIG) on student achievement. The SIG program receives substantial funding to act on improving student achievement via providing funding to lowest-achieving schools (those meet eligibility cutoff) with the condition that they implement one of the four intervention models; transformation, turnaround, restart, and closure. The eligibility cutoff is at the lowest fifth percentile of school level means on the academic achievement measures. Schools below the cutoff are assigned to the treatment (receives SIG) and remaining schools are denied participation. The study did not detect substantial effects on academic achievement measures, most of which are not statistically significant.

Sample Size and Bandwidth

Data was collected from 460 schools within 50 districts and 21 states before outcome variables were available. The sample size for each outcome across three outcome years is reported in Table 5 (Dragoset et al., 2019, p. 231) from which we can calculate the average number of students per school. In this illustration I focus on ex-ante MDES computations, therefore, bandwidth selection procedures are not applicable, however, we can compute MDES within ex-post bandwidth limits for which outcome measures are available. We use benchmark bandwidth values from Table 6 (Dragoset et al., 2019, p. 238).

This option can be incorporated into the MDES computations via altering truncation intervals. For example, for the first year math test score outcome, \( k_1 = \text{cutoff} - .86 \) and \( k_1 = \text{cutoff} + .86 \) where the cutoff is determined from the quantile of the standard normal distribution as \( \text{cutoff} = \text{qnorm}(p = .413) \) and where \( p \) is determined from the proportion of schools in treatment to all selected schools \( (p = 190 / 460) \). I assume \( p \) and cutoff does not change across various outcome measures and years (details are not available in the study).

Degrees of Freedom Adjustment for Fixed Effects

The authors report two distinct models for grades five and lower and grades six and higher samples. For pedagogical reasons, I only consider grades six and higher (K-6 to K-12), for which six indicator variables are included in the model as fixed effects. Besides, there are 20 indicator variables for state fixed effects, totaling 26 fixed effects without their interaction with the treatment indicator. There are also three school-level predictors that are thought to be correlated with the school level aggregate outcomes; averaged math and reading scores, and percentage of students eligible for free or reduced-price lunch. Moreover, I assume the score variable and its quadratic form is included in the model without interactions with the treatment (other variations are experimented and reported in Table 4). So in addition to the intercept and treatment indicator, there are 31 predictors for which regression coefficients are estimated. Since we have already taken into account the number of covariates at the cluster level, we do not need to change the default \( g^2 = 0 \). Therefore, degrees of freedom should be adjusted accordingly. For example, for 2010-2011 math test score outcome this adjustment is \( \text{df} = 390 - 33 \). Again, I assume this model applies to all outcome measures and years
(details are not available in the study), thus, while the number of clusters varies across outcome measures and years, 33 is subtracted from all.

**Intra-class Correlation Coefficient and Explanatory Power of Covariates**

The proportion of variance in the outcome explained by the predictors at different levels is not available in the study. Sun, Penner, and Loeb (2017) provides adjusted R-squared value around .65 for math and a little over .70 for English language arts (ELA) test scores, although they are based on the data pooled across outcome years and are based on student-level outcomes. Sun and colleagues included prior student achievement, student covariates, grade, and school fixed effects. Hedges and Hedberg (2013) provide a more detailed picture, both providing intra-class correlation coefficients and R-squared values at different levels based on pretest scores and minimal demographic variables. Since we focus on the sample from grade 6 to grade 12, we extract this information from Hedges and Hedberg (2013).

Intra-class correlation for grade 6 math test score is reported as .19, averaged across 11 states. Pretest explains .66 of the variation in level 1 outcome, and .58 of the variation in aggregate level 2 outcome. Minimal demographic information includes level 1 gender, race, socioeconomic status and English learner status, and their means introduced at level 2. Minimal demographic variables explain .57 of the variation in level 1 outcome, and .09 of the variation in aggregate level 2 outcome. As for the reading test scores, intra-class correlation for grade 6 is .17, averaged across 11 states. Pretest explains .76 of the variation in level 1 outcome and .55 of the variation in aggregate level 2 outcome. Minimal demographic variables explain .69 of the variation in level 1 outcome, and .11 of the variation in aggregate level 2 outcome. For higher grades these values do not change substantially, so we do not mention them here for the sake of brevity.

Finally, to be conservative, I specify the highest intra-class correlation across grades 6 to 12 for both reading and math with argument \( \rho_2 = .25 \). I assume the pretest score and minimal demographic variables explain 70% of the variance in the level 1 outcome, and aggregate pretest score and fixed effects explain 60% of the variance in the aggregate level 2 outcome with arguments \( r_{21} = .70 \) and \( r_{22} = .60 \).

**Lee and Card (2008) Specification Bias**

Dragoset et al. (2019) stratified schools within unique values of the score variable following Lee and Card (2008). Within the specified bandwidths, the clustering effect resulting from the unique values of the score variable is rather low. There are 98 schools within 95 unique values of the score variable for math in the year 2010-2011 and 2011-2012, and 115 schools within 112 unique values of the score variable in the year 2012-2013. Therefore, I avoid introducing another layer of clustering and continue with the two-level model.

**Multiple Comparisons**

There are four intervention groups but results are reported in aggregate, so I will assume there is one SIG intervention versus a control group, and 6 outcome measures (math
and reading only for pedagogical reasons) investigated across three outcome years. Both the smallest Benjamini and Hochberg (1995) p-values and Benfornni adjustment result in the same p-value when 6 separate tests are considered. Thus, this adjustment can be specified with the argument alpha = .0083.

**Treatment Group Participation and Control Group Crossovers**

The treatment group participation rate is 85%, and the control group crossover rate is 10%. This means while 85% of schools assigned to treatment implemented one of the four intervention models, only 10% of the control group crossed-over to the treatment group. Thus, fuzzy CRD was implemented, which produce local average treatment effect (LATE) estimates. This can be specified with arguments rate.tp = .85, and rate.cc = .10. I assume rates do not change across various outcome measures and years (details are not available in the study).

**Minimum Detectable Effect Sizes**

Effect sizes reported in the SIG effectiveness literature range from 0.10 to 1.11 for math, and 0.10 to 0.83 for reading mostly hovering around 0.20 per Table 8 (Dragoset et al., 2019, p. 242). Dragoset et al. (2019) report ex-ante MDES values ranging from 0.19 to 0.22. When multiple comparisons, treatment group participation, and control-group crossover rates are taken into consideration, MDES values for LATE estimates would be larger for a hypothetical future study. For quadratic functional form without interactions, largest MDES value across two subjects and three outcome years is 0.30, 95% CI [0.072, 0.525], and this does not change substantially for the linear functional form, with or without interactions. When interactions are included for the quadratic functional form, largest MDES value across two subjects and three outcome years is 0.45, 95% CI [0.107, 0.783]. Larger MDES values in comparison Dragoset et al. (2019) are mostly driven by Type I error adjustment for multiple comparisons and correction for fuzzy estimation. Interactions inflate MDES values with quadratic functional form and beyond, but not with the linear functional form due to symmetric truncation intervals.

**Conclusion and Discussion**

There are two common approaches to estimate treatment effects in multilevel RDs: (i) Based on cluster-adjusted non-parametric local polynomial estimation (NLPE), and (ii) parametric mixed model estimation (PMME). Bandwidth selection algorithms in NLPE achieve at least two things: First, the relationship with the outcome is more likely to be linear, favoring parsimony in the model construction. Second, resemblance to randomized control trials (RCTs) - in terms of estimate unbiasedness - is achieved in part owing to random measurement error mimicking random assignment mechanism (Boruch, 1975; Campbell & Stanley, 1963; Lee & Lemieux, 2010). However, the tradeoff is, external validity is weaker in comparison, power rates are reduced in general, and nominal Type I errors are inflated.

When outcome data is not available, presumptuous data collection may implicate the disposal of significant information away from the cutoff. It is for this reason that simpler
functional forms prevail with narrower bandwidths. NLPE reduces dependency on the functional form via focusing on a narrow range around the cutoff, thus, results may not suffer from ignorance at the hands of a naive analyst as much as it would with PMME. PMME perspective is a strong alternative when one is confident in the functional form. It has higher power rates while keeping the nominal Type I error rates.

From the design perspective, this study adopts the PMME approach and handles MDES computation via deriving closed-form formulas for sampling variance of the treatment effect in CRD studies. Schochet’s (2008) formulas are extended beyond linear functional form, and they are standardized in terms of commonly reported standardized variance parameters. An empirical framework based on a hybrid simulation approach (only RDDE section of the variance formula is simulated) is introduced in which functional forms beyond quadratic functional form or interactions with treatment indicator can be inspected.

There are several advantages of MDES computations based on PMME. It explicitly acknowledges the fact that the score variable is at the cluster level. Variance-reducing properties of the explanatory power of covariates at different levels of the hierarchy can be inspected, and optimal design can be performed when there are monetary constraints. There are also limitations to this approach. If a researcher uses this approach to compute ex-ante MDES for a study, but use NLPE ex-post, most likely it will be underpowered. Underspecified functional forms may produce consequential results, thus erring on the overspecified functional forms is more friendly in comparison.

Although economists have recently begun to engage in power analysis considering complex models, they mostly rely on data-driven approaches (Cattaneo, Titiunik, & Vazquez-Bare, 2019) producing sample-based estimates. This line of work certainly deserves attention; however, the prevailing notion among scholars in evaluation research is that design and analysis should be distinct. Moreover, from PMME framework perspective, scholars have the option to consider (or model) treatment effect heterogeneity though not applicable to this study, increase explanatory power of covariates, determine sample size at distinct levels and optimally devise a study when there are monetary constraints (Bulus & Dong, 2019; Bulus & Sahin, 2019).

When there are fixed-site effects, they can be incorporated into the unconditional and full models during the estimation of variance parameters ahead of power computations. Besides, degrees of freedom can be modified such that there are as many intercepts and as many treatment effects as sites (therefore two times as many sites are extracted from unmodified degrees of freedom). Otherwise, formulas for CRDs are retained with fixed site effects. There are other alternatives where only intercept is estimated for each site but the treatment effect is considered constant. This is merely extracting number of sites from unmodified degrees of freedom along with including fixed effects into the statistical models without their interaction with the treatment condition. The latter fixed effect adjustment in MDES computation was used in the illustrative example earlier.

While empirical framework provides flexibility, the question of whether we ever need polynomials beyond second-order comes to mind. We may not need higher order
polynomials, not even second-order itself with NPLE if there does not exist a compelling reason to practice otherwise. From the design perspective, RDDE remains the same with symmetric bandwidths whether we consider linear or quadratic functional forms, however, it jumps drastically when cubic form is considered. This also means, when the score variable in the true model follows cubic form, underspecification could be a grave mistake. The chances of this happening with NPLE are probably minuscule. It is more likely to happen with PMME when the full range of score distribution is considered. Including higher order polynomials in the estimation routines might be detrimental to RDD estimates (Gelman & Imbens, 2017; Gelman & Zelizer, 2015), nonetheless, underspecified functional form with parametric methods can produce consequential results as well.

MDES computations in CRDs considering quadratic functional form, and empirical extensions beyond might be an improvement over existing methods, yet there are other issues need to be investigated. Derivations and empirical framework can be extended to blocked (multisite) individual-level regression discontinuity studies (BIRD2, BIRD3, and BIRD4) and blocked cluster-level regression discontinuity studies (BCRD3r2, BCRD4r3) studies. Multiple score variables (Porter, Reardon, Unlu, Bloom, & Cimpian, 2016) and multiple cutoffs might exist (Cattaneo, Keele, Titiunik, & Vazquez-Bare, 2019), besides, a score variable may be random across blocks all of which warrant further investigation. Another line of research is to investigate MDES computations in light of moderator and mediator effects analogous to recent advancements in CRT literature (e.g., Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017; Spybrook, Kelcey, & Dong, 2016).

References

Andrews, I., & Oster, E. (2017). Weighting for external validity (No. w23826). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w23826.pdf

Angrist, J., & Imbens, G. (1995). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2), 467-475. doi:10.2307/2951620

Bertanha, M., & Imbens, G. W. (2019): External Validity in Fuzzy Regression Discontinuity Designs, Journal of Business & Economic Statistics. doi:10.1080/07350015.2018.1546590

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300. Retrieved from http://www.jstor.org/stable/2346101

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547-556. doi:10.1177/0193841X9501900504

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. MDRC Working Papers on Research Methodology. New York, NY: MDRC. Retrieved from https://files.eric.ed.gov/fulltext/ED493363.pdf
Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59. doi:10.3102%2F0162373707299550

Boruch, R. F. (1975). Coupling Randomized Experiments and Approximations to Experiments in Social Program Evaluation. *Sociological Methods & Research, 4*(1), 31-53. doi:10.1177/004912417500400103

Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-School Variation in Students' Achievement, Motivation, Affect, and Learning Strategies: Results from 81 Countries for Planning Group-Randomized Trials in Education. *Journal of Research on Educational Effectiveness, 11*(3), 452-478. doi:10.1080/19345747.2017.1375584

Bulus, M., & Dong, N. (2019a). 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*. Advance online publication. doi:10.1080/00220973.2019.1636197

Bulus, M., & Dong, N. (2019b). cosa: Bound Constrained Optimal Sample Size Allocation. R package version 2.1.0. [Software]. URL: https://cran.r-project.org/package=cosa

Bulus, M., & Sahin, S. G. (2019). Estimation and Standardization of Variance Parameters for Planning Two- and Three-level Cluster Randomized Trials: A Short Guide for Researchers. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 179-201. doi:10.21031/epod.530642

Burkardt, J. (2014). The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*. Retrieved from https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2019). Regression Discontinuity Designs Using Covariates. *The Review of Economics and Statistics, 101*(3), 442-451. doi:10.1162/rest_a_00760

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2018). rdrobust: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs. R package version 0.99.4. https://cran.r-project.org/package=rdrobust

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2018). Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs (v2). Retrieved from https://arxiv.org/abs/1809.00236

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N.L. Gage (Ed.), *Handbook of Research on Teaching* (pp. 1-76). Chicago, IL: Rand McNally.

Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2019). Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs. arXiv preprint arXiv:1808.04416. https://arxiv.org/pdf/1808.04416.pdf
Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2019). Power calculations for regression-discontinuity designs. *The Stata Journal, 19*(1), 210-245. doi:10.1177/2193841X19830919

Cerulli, G., Dong, Y., Lewbel, A., & Poulsen, A. (2017). Testing stability of regression discontinuity models. In *Regression Discontinuity Designs: Theory and Applications* (pp. 317-339). Emerald Publishing Limited. Retrieved from http://www.yingyingdong.com/Research/RD-Stability-AIE2017.pdf

Cook, T. D. (2008). “Waiting for Life to Arrive”: A history of the regression discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics, 142*(2), 636-654. doi:10.1016/j.jeconom.2007.05.002

Deke, J., & Dragoset, L. (2012). *Statistical Power for Regression Discontinuity Designs in Education: Empirical Estimates of Design Effects Relative to Randomized Controlled Trials*. Princeton, NJ: Mathematica Policy Research. Retrieved from https://files.eric.ed.gov/fulltext/ED533141.pdf

Deke, J., Wei, T., Kautz, T. (2017). *Asymdystopia: The threat of small biases in evaluations of education interventions that need to be powered to detect small impacts*. (NCEE 2018–4002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from https://files.eric.ed.gov/fulltext/ED576329.pdf

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful Effect Sizes, Intraclass Correlations, and Proportions of Variance Explained by Covariates for Planning Two- and Three-Level Cluster Randomized Trials of Social and Behavioral Outcomes. *Evaluation Review, 40*(4), 334-377. doi:10.1177/0193841X16671283

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., & Lee Luca, D. (2019). The Impact of School Improvement Grants on Student Outcomes: Findings From a National Evaluation Using a Regression Discontinuity Design. *Journal of Research on Educational Effectiveness, 12*(2), 215-250. doi:10.1080/19345747.2019.1571654

Dimmery, D. (2016). rdd: Regression Discontinuity Estimation. R package version 0.57. https://cran.r-project.org/package=rdd

Gelman, A., & Imbens, G. (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics, 37*(3), 447-456. doi:10.1080/07350015.2017.1366909

Gelman, A., & Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research & Politics, 2*(1), 1-7. doi:10.1177%2F2053168015569830

Goldberger, A. S. (1972a). Selection bias in evaluating treatment effects: Some formal illustrations. In T. Fomby, R. C. Hill, D. L. Millimet, J. A. Smith, E. J. Vytlacil (Eds.). *Modeling and Evaluating Treatment Effects in Econometrics* (pp. 1-31). Amsterdam, Netherlands: JAI Press. doi:10.1016/S0731-9053(07)00001-1

Goldberger, A. S. (1972b). *Selection bias in evaluating treatment effects: The case of...
Juras, R. (2016). Estimates of Intraclass Correlation Coefficients and Other Design Parameters for Studies of School-Based Nutritional Interventions. Evaluation Review, 40(4), 314-333. doi:10.1177/0193841X16675223

Kelcey B, Dong N, Spybrook J, Cox K (2017). Statistical Power for Causally Defined Indirect Effects in Group-Randomized Trials With Individual-Level Mediators. Journal of Educational and Behavioral Statistics, 42(5), 499–530. doi:10.3102/1076998617695506

Kelcey B, Dong N, Spybrook J, Shen Z (2017). Experimental power for indirect effects in group-randomized studies with group-level mediators. Multivariate Behavioral Research, 52(6), 699–719. doi:10.1080/00273171.2017.1356212

Konstantopoulos, S. (2009). Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. Evaluation Review, 33(4), 335-357. doi:10.1177/0193841X09337991

Konstantopoulos, S. (2011). Optimal Sampling of Units in Three-Level Cluster Randomized Designs: An ANCOVA Framework. Educational and Psychological Measurement, 71(5), 798-813. doi:10.1177/0013164410397186
Konstantopoulos, S. (2013). Optimal Design in Three-Level Block Randomized Designs with Two Levels of Nesting: An ANOVA Framework with Random Effects. *Educational and Psychological Measurement*, 73(5), 784-802. doi:10.1177/0013164413485752

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2), 655-674. doi:10.1016/j.jeconom.2007.05.003

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2), 281-355. doi:10.1257/jel.48.2.281

Liu, X. (2003). Statistical Power and Optimum Sample Allocation Ratio for Treatment and Control Having Unequal Costs per Unit of Randomization. *Journal of Educational and Behavioral Statistics*, 28(3), 231-248. doi:10.3102/10769986028003231

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698-714. doi:10.1016/j.jeconom.2007.05.005

Moerbeek, M., & Safarkhani, M. (2018). The Design of Cluster Randomized Trials With Random Cross-Classifications. *Journal of Educational and Behavioral Statistics*, 43(2), 159-181. doi:10.3102/1076998617730303

Porter, K. E., Reardon, S. F., Unlu, F., Bloom, H. S., & Cimpian, J. R. (2017). Estimating Causal Effects of Education Interventions Using a Two-Rating Regression Discontinuity Design: Lessons From a Simulation Study and an Application. *Journal of Research on Educational Effectiveness*, 10(1), 138-167. doi:10.1080/19345747.2016.1219436

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. Psychological Methods, 2(2), 173. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Liu, X. (2000). Statistical Power and Optimal Design for Multisite Trials. Psychological Methods, 5, 199-213. doi:10.1037/1082-989X.5.2.199

Rhoads, C. H., & Dye, C. (2016). Optimal Design for Two-Level Random Assignment and Regression Discontinuity Studies. *The Journal of Experimental Education*, 84(3), 421-448. doi:10.1080/00220973.2015.1048846

Rickles, J., Zeiser, K., & West, B. (2018). Accounting for Student Attrition in Power Calculations: Benchmarks and Guidance. *Journal of Research on Educational Effectiveness*, 11(4), 622-644. doi:10.1080/19345747.2018.1502384

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5(2), 121-125. doi:10.1007/BF00143942

Schochet, P. Z. (2008). *Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations* (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from 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. Retrieved from https://www.jstor.org/stable/40263528

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two- and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41*(6), 605-627. doi:10.3102/1076998616655442

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open, 2*(1). doi:10.1177/2332858415625975

Stigler, M., & Quast, B. (2015). rddtools: Toolbox for Regression Discontinuity Design (‘RDD’). R package version 0.4.0. https://cran.r-project.org/package=rddtools

Sun, M., Penner, E. K., & Loeb, S. (2017). Resource- and approach-driven multidimensional change: Three-year effects of school improvement grants. *American Educational Research Journal, 54*(4), 607–643. doi:10.3102/002831217695790

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review, 37*(6), 490-519. doi:10.1177/0193841X14531584

Wing, C., & Bello-Gomez, R. A. (2018). Regression discontinuity and beyond: Options for studying external validity in an internally valid design. *American Journal of Evaluation, 39*(1), 91-108. doi:10.1177%2F1098214017736155
### Tables and Figures

#### Table 1. Minimum Required Number of Clusters for Two-Level CRD

| $(k_1, k_2)$ | $\rho_{T2}$ | $|\rho_{T2}|$ | $\rho_{Z2}$ | Linear | Quadratic | $\rho_{T2}$ | $|\rho_{T2}|$ | $\rho_{Z2}$ | Linear | Quadratic |
|-------------|-------------|---------------|-------------|--------|-----------|-------------|---------------|-------------|--------|-----------|
| (-1, 1)     | 0.54        | 0.56          | 0           | 846    | 1494      | 0.52        | 0.54          | 0           | 824    | 1359      |
| (-2, 2)     | 0.57        | 0.57          | 0           | 893    | 1731      | 0.52        | 0.54          | 0           | 824    | 1359      |
| (-3, 3)     | 0.58        | 0.55          | 0           | 913    | 1665      | 0.52        | 0.54          | 0           | 824    | 1359      |
| (-1, 2) or (-2, 1) | 0.61      | 0.84          | 0.66        | 963    | 2049      | 0.52        | 0.74          | 0.79        | 824    | 1359      |
| (-2, 1) or (-1, 2) | 0.50      | 0.07          | -0.66       | 799    | 1272      | 0.52        | 0.08          | -0.79       | 824    | 1359      |
| (-2, 3) or (-3, 2) | 0.60      | 0.67          | -0.22       | 943    | 1787      | 0.52        | 0.74          | 0.61        | 824    | 1359      |
| (-3, 2) or (-2, 3) | 0.55      | 0.42          | -0.22       | 867    | 1561      | 0.52        | 0.11          | -0.61       | 824    | 1359      |
| (-1, 3) or (-3, 1) | 0.65      | 0.83          | 0.72        | 1033   | 1948      | 0.52        | 0.71          | 0.89        | 824    | 1359      |
| (-3, 1) or (-1, 3) | 0.48      | 0.01          | -0.72       | 779    | 1153      | 0.52        | -0.22         | -0.89       | 824    | 1359      |

#### Note.
Results assume that there are 20 level 1 unit on average per level 2 unit, intra-class correlation coefficient for level 2 is 0.40, a two-tailed hypothesis test is conducted for an effect size of 0.25 with Type I error rate of 0.05 and power rate of 80%. 

---

**MDES IN CLUSTERED RDD: QUADRATIC FUNCTIONAL FORM AND BEYOND**

27
Table 2. Minimum Required Number of Clusters for Three-Level CRD

| (k₁, k₂) | Truncated Normal (μ = 0, σ = 1, a = k₁, b = k₂) | Uniform (a = k₁, b = k₂) |
|----------|-----------------------------------------------|-------------------------|
|          | Correlations | Number of Clusters | Correlations | Number of Clusters |
|          | ρ_{TZ} | ρ_{T2}^2 | Linear | Quadratic | ρ_{TZ} | ρ_{T2}^2 | Linear | Quadratic |
| (-1, 1)  | 0.54   | 0.56    | 0      | 669      | 1.18   | 0.54    | 0      | 652      |
| (-2, 2)  | 0.57   | 0.57    | 0      | 706      | 1.36   | 0.54    | 0      | 652      |
| (-3, 3)  | 0.58   | 0.55    | 0      | 722      | 1.31   | 0.54    | 0      | 652      |
| (-1, 2)  | 0.60   | 0.80    | 0.66   | 762      | 1.62   | 0.52    | 0.74   | 0.79     | 652      | 1.075 |
| (-2, 1)  | 0.50   | 0.70    | -0.66  | 632      | 1.06   | 0.52    | 0.08   | -0.79    | 652      | 1.075 |
| (-3, 2)  | 0.60   | 0.67    | 0.22   | 746      | 1.41   | 0.52    | 0.74   | 0.61     | 652      | 1.075 |
| (-3, 3)  | 0.58   | 0.42    | -0.22  | 686      | 1.23   | 0.52    | 0.11   | -0.61    | 652      | 1.075 |
| (-1, 3)  | 0.65   | 0.22    | 0.72   | 817      | 1.54   | 0.52    | 0.71   | 0.89     | 652      | 1.075 |
| (-3, 1)  | 0.65   | 0.11    | -0.72  | 617      | 0.91   | 0.52    | -0.22  | -0.89    | 652      | 1.075 |

Note. Results assume that there are 20 level 1 units on average per level 2 unit, and 3 level 2 units on average per level 3 unit, intra-class correlation coefficient for level 2 and level 3 are .40 and .20, respectively, a two-tailed hypothesis test is conducted for an effect size of 0.25 with Type I error rate of 5% and power rate of 80%.
Figure 1. Correlations between $T$, $Z$, $Z^2$ triad ($\rho_{Z^2Z^2} = 0$), and RDDE as a function of $p$ and distribution type.
Figure 2. Correlations between $T, Z, Z^2$ triad ($\rho_{Z^2} = 0$), and RDDE as function of $p$ and truncation interval.
Figure 3. Correlations between $T$, $Z$, $Z^2$ triad and RDDE as a function of $p$ and skewness. Figures are based on empirical skewed distribution ($N = 10000$), and values in paranthesis are based on sample size adjusted Fisher-Pearson coefficient of skewness.
Figure 4. RDDE as a function of polynomial order, interactions with treatment, and $p$. Lines for linear and quadratic form are based on derivations. Intermittent jittery lines for interactions and continuous but jittery line for the cubic form are based on simulations (1000 draws from the standard normal distribution averaged over 1000 replications).
Supplement - Statistical Models and Derivations for CRD3 and CRD4 Designs

Cluster-level Regression Discontinuity for Three Level Design (CRD3)

Unconditional Model for CRD3 Design

The following unconditional model is used to obtain variance parameters $\sigma^2$ and $\tau^2$ as defined below, which will be used to calculate intra-class correlation coefficients and R-squared values along with parameters from the full model.

Level 1: $Y_{ijk} = \beta_{0jk} + r_{ijk}$
Level 2: $\beta_{0jk} = y_{00k} + \mu_{0jk}$
Level 3: $y_{00k} = \xi_{000} + \varsigma_{00k}$

where $r_{ijk} \sim N(0, \sigma^2)$, $\mu_{0jk} \sim N(0, \tau^2_Z)$, $\varsigma_{00k} \sim N(0, \tau^2_\varsigma)$, $\sigma^2$, $\tau^2_Z$ and $\tau^2_\varsigma$ are unconditional variances in the outcome between level 1, level 2 and level 3 units, respectively.

Full Model for CRD3 Design

The following model is used to obtain variance parameters $\sigma^2_{X}$, $\tau^2_{W}$ and $\tau^2_{T}$, as defined below, which are used to obtain R-squared values along with the parameters from the unconditional model.

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{ijk} + r_{ijk}$
Level 2: $\beta_{0jk} = y_{00k} + y_{01k}W_{jk} + \mu_{0jk}$
$\beta_{1jk} = y_{10k}$
Level 3: $y_{00k} = \xi_{000} + \xi_{001}T_k + \xi_{002}(Z_k - Z_0) + \xi_{003}(Z_k - Z_0)^2 + \xi_{004}V_k + \varsigma_{00k}$
$y_{01k} = \xi_{010}$
$y_{10k} = \xi_{100}$

where $r_{ijk} \sim N(0, \sigma^2_{X})$, $\mu_{0jk} \sim N(0, \tau^2_{W})$, $\varsigma_{00k} \sim N(0, \tau^2_{T})$, $\rho_2 = \tau^2_Z / (\tau^2_\varsigma + \tau^2_Z + \sigma^2)$ and represents proportion of variance in the outcome between level 2 units, $\rho_3 = \tau^2_{W} / (\tau^2_Z + \tau^2_\varsigma + \sigma^2)$ and represents proportion of variance in the outcome between level 3 units, $\sigma^2$, $\tau^2_Z$, and $\tau^2_\varsigma$ are level 1, level 2 and level 3 residual variances in the unconditional model, respectively, $\sigma^2_{X}$, $\tau^2_{W}$ and $\tau^2_{T}$ are level 1, level 2 and level 3 variances conditional on predictors at their respective level, $R^2_1 = 1 - \sigma^2_{X} / \sigma^2$ and is level 1 variance explained by level 1 predictors, $R^2_2 = 1 - \tau^2_Z / \tau^2_\varsigma$ and is proportion of variance at level 2 explained by level 2 predictors, $R^2_3 = 1 - \tau^2_{T} / \tau^2_3$ and is proportion of variance at level 3 explained by level 3 predictors.

Derivation of Asymptotic Variance for CRD3 Design

The generalized least square estimate of variance – covariance matrix for fixed effect coefficients can be stated as
where $X_k$ is $Jn\times 4$ design matrix for cluster $k$ including intercept, $T_k$, $Z_k$, $Z_k^2$, $V = ZGZ^T + R$, $R$ is $(nJK)\times(nJK)$ diagonal level 1 residual variance matrix, $G$ is $(JK + K)\times(JK + K)$ diagonal level 2 and level 3 residual variance matrix, $Z$ is $(nJK)\times(JKx1 + Kx1)$ matrix of random effects (note that $x1$ is due having only intercept random across level 2 and level 3 units) and can be obtained via $Z = J^T \otimes X^T$ where $J$ is indicator matrix of group membership for each observation and $X$ is design matrix for random effects. The resultant $V$ matrix is $(nJK)\times(nJK)$ diagonal level 2 and level 3 residual variance matrix, $G$ is $(JK + K)\times(JK + K)$ diagonal level 2 and level 3 residual variance matrix, $Z$ is $(nJK)\times(JKx1 + Kx1)$ matrix of random effects (note that $x1$ is due having only intercept random across level 2 and level 3 units) and can be obtained via $Z = J^T \otimes X^T$ where $J$ is indicator matrix of group membership for each observation and $X$ is design matrix for random effects. The resultant $V$ matrix is $(nJK)\times(nJK)$ a nested block diagonal sparse matrix where each level 3 block diagonal structure represent the total variance. For example assume $n = 2$, $J = 2$, and $K = 2$ then diagonal elements of $V$ matrix for unconditional model takes the form

$$V_k = \begin{bmatrix}
\sigma^2 + \tau_2^2 + \tau_3^2 & \tau_2^2 + \tau_3^2 & \tau_3^2 \\
\tau_2^2 + \tau_3^2 & \sigma^2 + \tau_2^2 + \tau_3^2 & \tau_3^2 \\
\tau_3^2 & \tau_3^2 & \sigma^2 + \tau_2^2 + \tau_3^2
\end{bmatrix}$$

Assuming constant variance and compound symmetry structure, only intercept is random across level 2 so

$$cov(\xi) = \frac{(Jn\tau_{31|\tau,Z,Z^2,V}^2 + n\tau_{21|W}^2 + \sigma_X^2)}{K}(X^TX)^{-1}$$

The cell associated with the treatment effect is

$$cov(\xi)_{[2,2]} \xrightarrow{p} \frac{(Jn\tau_{31|\tau,Z,Z^2,V}^2 + n\tau_{21|W}^2 + \sigma_X^2)}{KJnP(1-P)}(RDDE)$$

Based on the parameters defined above this formula can be standardized as

$$var(\xi_{001}) = \frac{Jn\rho_3(1 - R_2^2) + n\rho_2(1 - R_2^2) + (1 - \rho_3 - \rho_2)(1 - R_2^2)}{KJnP(1-P)}(RDDE)$$

where the treatment effect is reparametrized in terms of Cohen’s $d$ as $\xi_{001}^* = \xi_{001}/\sqrt{\sigma^2 + \tau_2^2 + \tau_3^2},$ and

$$RDDE = \frac{1 - \rho_{ZZ^2}^2}{1 - \rho_{TZ}^2 - \rho_{TZ^2}^2 - \rho_{ZZ^2}^2 + 2\rho_{TZ}\rho_{TZZ}\rho_{ZZ^2}}$$

Cluster-level Regression Discontinuity for Four Level Design (CRD4)

Unconditional Model for CRD4 Design
The following unconditional model is used to obtain variance parameters $\sigma^2$ and $\tau^2$ as defined below, which will be used to calculate intra-class correlation coefficients and R-squared values along with parameters from the full model.

Level 1: $Y_{ijkl} = \beta_{0ijkl} + \eta_{ijkl}$
Level 2: $\beta_{ijkl} = \gamma_{00kkl} + \mu_{0ijkl}$
Level 3: $\gamma_{00kkl} = \xi_{000kl} + \zeta_{000kl}$
Level 4: $\xi_{0000l} = \zeta_{00000} + \upsilon_{0000l}$

where $\eta_{ijkl} \sim N(0, \sigma^2), \mu_{0ijkl} \sim N(0, \tau^2_2), \zeta_{000kl} \sim N(0, \tau^2_3), \upsilon_{0000l} \sim N(0, \tau^2_4)$ $\sigma^2$, $\tau^2_2$, $\tau^2_3$ and $\tau^2_4$ are unconditional variances in the outcome between level 1, level 2, level 3 and level 4 units, respectively.

**Full Model for CRD4 Design**

The following model is used to obtain variance parameters $\sigma^2_{X|W}$, $\tau^2_{Z|W}$, and $\tau^2_{Z|V}$, as defined below, which are used to obtain R-squared values along with the parameters from the unconditional model.

Level 1: $Y_{ijkl} = \beta_{0ijkl} + \beta_{1ijkl}X_{ijk} + \eta_{ijkl}$
Level 2: $\beta_{1ijkl} = \gamma_{00kkl} + \gamma_{01kkl}W_{ijkl} + \mu_{0ijkl}$
Level 3: $\gamma_{00kkl} = \xi_{000kl} + \xi_{001kl}V_{kl} + \zeta_{000kl}$
Level 4: $\xi_{0000l} = \zeta_{00000} + \zeta_{0001l}T_k + \zeta_{0002l}(Z_k - Z_0) + \zeta_{0003l}(Z_k - Z_0)^2 + \zeta_{0004l}Q_l + \upsilon_{0000l}$

$\xi_{0010l} = \zeta_{0010l}$

where $\eta_{ijkl} \sim N(0, \sigma^2_{X|W}), \mu_{0ijkl} \sim N(0, \tau^2_{Z|W}), \zeta_{000kl} \sim N(0, \tau^2_3|V), \upsilon_{0000l} \sim N(0, \tau^2_{Z|V})$, $\rho_2 = \tau^2_2/\left(\tau^2_2 + \tau^2_3 + \tau^2_4 + \sigma^2\right)$ and represents proportion of variance in the outcome between level 2 units, $\rho_3 = \tau^2_3/\left(\tau^2_2 + \tau^2_3 + \tau^2_4 + \sigma^2\right)$ and represents proportion of variance in the outcome between level 3 units, $\rho_4 = \tau^2_4/\left(\tau^2_2 + \tau^2_3 + \tau^2_4 + \sigma^2\right)$ and represents proportion of variance in the outcome between level 4 units, $\sigma^2$, $\tau^2_2$, $\tau^2_3$, and $\tau^2_4$ are level 1, level 2, level 3 and level 4 residual variances in the unconditional model, respectively, $\sigma^2_{X|W}$, $\tau^2_{Z|W}$, $\tau^2_{Z|V}$ and $\tau^2_{4|Z,Q}$ are level 1, level 2 and level 3 variances conditional on predictors at their respective level, $R^2_1 = 1 - \sigma^2_{X|W}/\sigma^2$ and is level 1 variance explained by level 1 predictors, $R^2_2 = 1 - \tau^2_{Z|W}/\tau^2_2$ and is proportion of variance at level 2 explained by level 2 predictors, $R^2_3 = 1 - \tau^2_{3|V}/\tau^2_3$ and is proportion of variance at level 3 explained by level 3 predictors, $R^2_4 = 1 - \tau^2_{4|Z,Q}/\tau^2_4$ and is proportion of variance at level 4 explained by level 3 predictors.

**Derivation of Asymptotic Variance for CRD4 Design**
The generalized least square estimate of variance – covariance matrix for fixed effect coefficients can be stated as

$$cov(\zeta) = \left( \sum_{l=1}^{L} X_l^T V_l^{-1} X_l \right)^{-1}$$

where $X_l$ is $KJn\times 4$ design matrix for cluster $l$ including intercept, $T_l$, $Z_l$, $Z_l^2$, $V = ZGZ^T + R$. $R$ is $(nJKL)\times(nJKL)$ diagonal level 1 residual variance matrix, $G$ is $(JKL + JK + L)\times(JKL + JK + L)$ diagonal level 2, level 3 and level 4 residual variance matrix, $Z$ is $(nJKL)\times (JKLx1 + KLex1 + Lx1)$ matrix of random effects (note that $x1$ is due to having only intercept random across level 2, level 3, and level 4 units) and can be obtained via $Z = JT \otimes X^T$ where $J$ is indicator matrix of group membership for each observation and $X$ is design matrix for random effects. Because only intercept is random $Z = J^T$. The resultant $V$ matrix is $(nJKL)\times(nJKL)$ a nested block diagonal sparse matrix where diagonal structure represent the total variance. For example assume $n = 2$, $J = 2$, $K = 2$, and $L = 2$ then diagonal elements of $V$ matrix for unconditional model takes the form
MDES IN CLUSTERED RDD: QUADRATIC FUNCTIONAL FORM AND BEYOND

\[ V_i = \begin{bmatrix}
\sigma^2 + \tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_1^2 + \tau_2^2 & \tau_1^2 + \tau_2^2 & \tau_1^2 + \tau_2^2 & \tau_1^2 + \tau_2^2 & \tau_1^2 + \tau_2^2 \\
\tau_1^2 + \tau_2^2 + \tau_3^2 & \sigma^2 + \tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 \\
\tau_1^2 + \tau_2^2 + \tau_3^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \sigma^2 + \tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 \\
\tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \sigma^2 + \tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 \\
\tau_1^2 + \tau_2^2 + \tau_3^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 \\
\tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 & \tau_2^2 + \tau_3^2 + \tau_4^2 \\
\end{bmatrix} \]
Assuming constant variance and compound symmetry structure, only intercept is random across level 2, level 3, and level 4 so

$$\text{cov}(\xi) = \frac{(KJn\tau_4^2 + Jn\tau_3^2 + n\tau_2^2 + \sigma^2)}{L}(X^TX)^{-1}$$

The cell associated with the treatment effect is

$$\text{cov}([2], [2]) \sim p\left(\frac{(KJn\tau_4^2 + Jn\tau_3^2 + n\tau_2^2 + \sigma^2)}{LKJnP(1 - P)}(RDDE)\right)$$

Based on the parameters defined above this formula can be standardized as

$$\text{var}(\xi_{0001}) = \frac{KJn\rho_4(1 - R_4^2) + Jn\rho_3(1 - R_3^2) + n\rho_2(1 - R_2^2) + (1 - \rho_4 - \rho_3 - \rho_2)(1 - R_1^2)}{LKJnP(1 - P)}(RDDE)$$

where the treatment effect is reparametrized in terms of Cohen’s $d$ as $\xi_{0001} = \frac{\xi_{0001}}{\sqrt{\sigma^2 + \tau_2^2 + \tau_3^2 + \tau_4^2}}$, and

$$RDDE = \frac{1 - \rho_{zz}^2}{1 - \rho_{TZ}^2 - \rho_{Tz}^2 - \rho_{ZZ}^2 + 2\rho_{TZ}\rho_{Tz}\rho_{ZZ}}$$