

# Multiple-cutoff regression discontinuity designs in educational evaluation

HyeonJin Yoon<sup>1</sup>, Keith Zvoch<sup>2</sup>, Keith Smolkowski<sup>3</sup>, Ben Clarke<sup>4</sup>

<sup>1</sup>Nebraska Center for Research on Children, Youth, Families, and Schools, University of Nebraska-Lincoln, Lincoln, United States

<sup>2</sup>Department of Educational Studies, College of Education, University of Oregon, Eugene, United States

<sup>3</sup>Oregon Research Institute, Eugene, United States

<sup>4</sup>Department of School Psychology, College of Education, Eugene, United States

## Article Info

### Article history:

Received May 5, 2024

Revised Feb 7, 2025

Accepted Feb 14, 2025

### Keywords:

Early math intervention  
Heterogeneity of treatment effect  
Multiple-cutoffs  
Multi-site evaluation  
Regression discontinuity designs

## ABSTRACT

The regression discontinuity design (RDD) can be used for evaluating cut-score based educational interventions. The design enables researchers and evaluators to estimate the local causal effect of an intervention provided to those who are most at-risk. In certain educational evaluations, an RDD with multiple cutoffs can also be applied, where different cutoffs across classrooms or schools are used to assign students to the intervention condition. The availability of multiple cutoffs allows estimation of a pooled average treatment effect across cutoffs and/or individual estimates at each cutoff location, allowing for the possibility of heterogeneous treatment effects. The purpose of this paper is to demonstrate the design and analysis of the multiple-cutoff regression discontinuity (MCRD) design for the evaluation of need-based educational programs. Using data obtained from a kindergarten mathematics intervention for low achieving students, we illustrate the utility of the MCRD design for examining the average and potential variability of the regression discontinuity (RD) treatment effect. Implications for evaluation research and practice in education are discussed.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

HyeonJin Yoon

Nebraska Center for Research on Children, Youth, Families, and Schools, University of Nebraska-Lincoln  
61 Carolyn Pope Edwards Hall Lincoln, Nebraska, 68588-0235, United States

Email: hyoon5@unl.edu

## 1. INTRODUCTION

Although the randomized control trial (RCT) provides the most straightforward route to causal inference, practical and ethical concerns can make its implementation in educational evaluation contexts challenging [1], [2]. For example, when an intervention program or service is offered only to students most in need, the mechanism by which students are assigned to conditions is predetermined by program administrators. However, if service eligibility is identified solely on the basis of a cutscore on preprogram measure (e.g., income level, reading score), researchers and evaluators may be able to productively leverage the assignment mechanism that is used to systematically allocate program resources. Cutscore-based assignment yields a regression discontinuity (RD) design that facilitates causal inference as the process of assignment to treatment conditions is completely known and can be accounted for in the impact model [3].

The RD design has become an increasingly popular alternative for supporting causal inference in program evaluation and other field-based research contexts [4], [5]. The attractiveness of RD stems from its relatively simple structure and fit with the need-based service delivery practices that are common in applied settings. The structure is simple as only an assignment variable that completely determines the allocation of individuals to conditions (i.e., those scoring on one side of a cutoff score receive treatment and those on the

other side do not) and scores on an outcome of interest are needed. Although the nonrandom distribution of individuals or units to conditions would seem to invalidate the design, incorporating the assignment variable and the treatment indicator in the analysis of program outcomes establishes conditional independence between treatment assignment and outcomes [6]. The key idea is that individuals just below and just above the cutoff are assumed to be similar in all respects (e.g., demographic characteristics), except for whether they received the treatment. Therefore, any difference in outcomes between the two groups can be attributed to the treatment [3], [6]. Despite this practical advantage, the basic RD design has several limitations that can reduce its utility. The basic RD design is less efficient than a comparable RCT due to the high correlation between the assignment variable and the binary treatment indicator, thus requiring more cases for the same statistical power [7]. Additionally, in the basic RD design, causal inference is limited to the cutoff as treatment and control cases do not overlap beyond the cutoff value [8]. Consequently, for those interested in assessing program effectiveness at other points in the assignment variable distribution, the local impact estimate may have somewhat limited value.

The basic RD design has evolved into multiple design variants to address the complexity of evaluation contexts in the field of education. One common variation is the utilization of multiple cutoffs in RD design. The multiple-cutoff regression discontinuity (MCRD) design arises from variations in the cutoffs used to assign individuals or units to conditions [9]–[12]. For example, school districts may use different reading assessment score cutoffs as the eligibility criteria for inviting students to summer school, or universities may use different income eligibility thresholds for determining financial aid recipients.

To date, the most common analytic method for the MCRD design is estimating an average RD treatment effect by pooling treatment estimates across cutoffs (i.e., the centering-and-polling approach) [9], [10]. The pooled RD estimate provides a summary of the average treatment effect across cutoffs. However, the use of the pooling method may ignore heterogeneity in the treatment effect estimates at different cutoff locations [9]. When treatment effects vary across cutoffs, simply averaging these effects can obscure evidence that the intervention was more or less effective for participants at different cut points.

As a result, the purpose of this paper is to demonstrate modeling approaches that allow for simultaneous estimation of the average MCRD treatment effect and the potential heterogeneity in effects across intervention sites. Using data from a kindergarten mathematics intervention targeting low achieving students at-risk for developing a learning disability, we illustrate the application of MCRD using parametric and nonparametric modeling [3], [13] and estimate and model intervention effect heterogeneity across cutoffs. The current demonstration utilizes small within-cutoff unit samples, which mirror real-world evaluation contexts well [14]. However, small samples reduce the statistical power of the RD estimates at each local cutoff [7]. Therefore, while we use both parametric and nonparametric methods to demonstrate the MCRD design and analysis, we only implement the parametric method to explore heterogeneity in impacts across treatment sites as the nonparametric method does not support the more complex model required for such analysis. The proposed method offers a practical solution for evaluating educational programs that use different cutoffs across multiple intervention units or sites, especially in contexts with small sample sizes, and provides a valuable contribution to research on evaluation methods in education.

## 2. METHOD

### 2.1. Data source

Mathematics intervention data were collected from 2012 to 2015 in conjunction with a university–school district collaboration designed to evaluate the impact of a small-group, tier 2 [15] kindergarten mathematics intervention, ROOTS. In the first two years of the project, study participants attended kindergarten in one of four school districts in Oregon, United States: one school district located in the Portland metropolitan area and three located in suburban and rural areas of western Oregon (51 classrooms from 13 schools in total). In years two and three, two school districts from the metropolitan area of Boston, Massachusetts, United States, also participated (55 classrooms from 9 schools in total).

### 2.2. ROOTS intervention

ROOTS is a 50-lesson, tier 2 kindergarten mathematics intervention targeting students with or at-risk for math learning disabilities. Direct math instruction was delivered to students in small groups (maximum of five students) across 106 classrooms. Each session lasted 20 minutes and took place five days a week over a 10-week period. ROOTS uses evidence-based practices to enhance number sense and whole-number concepts, covering counting and cardinality, number operations, and base ten/place value. The intervention incorporates four validated, explicit, and systematic instructional practices: teacher modeling, deliberate practice, visual representations of mathematics, and academic feedback. Additionally, ROOTS aims to promote student engagement by providing structured opportunities for student responses.

## 2.3. Measures

### 2.3.1. Assessing student proficiency in early number sense (ASPENS)

Assessing student proficiency in early number sense (ASPENS) was one of two screening measures that comprised the assignment variable used to identify students for inclusion in the ROOTS intervention. ASPENS was designed to screen and monitor the progress of students who are at risk for mathematics difficulty in Grades K through 1 [16]. The kindergarten ASPENS assessment comprises three curriculum-based measures: number identification (112 items), magnitude comparison (56 items), and missing numbers (42 items), each individually administered for one minute. Children name numbers quickly, identify greater numbers, and determine missing numbers from strings. Each subtest scores correct or incorrect responses, with a composite score derived from weighted subtest scores based on standard deviation (SD) of the raw scores [16]. For the current data, the pretest ASPENS composite ranged from 0 to 253.9 ( $M=57.01$ ,  $SD=47.68$ ), with acceptable test–retest reliabilities ( $r=.74-.85$ ).

### 2.3.2. Number sense brief (NSB)

The number sense brief (NSB), the second screening measure used to construct the assignment variable [17], identifies students at risk for later difficulties in advanced mathematics concepts. Students' numerical competencies, including counting knowledge, number recognition, comparisons, non-verbal calculation, and story problems are assessed. Administered individually and untimed, it comprises 33 possible items scored as correct or incorrect, yielding a total score range of 0 to 33. Test–retest reliabilities ranged from .78 to .86 [18]. Predictive validity, assessed by the correlation between NSB scores in Grades K through 1 and Grade 3 Woodcock–Johnson Achievement Test scores, ranged from .62 to .64 [18].

### 2.3.3. Stanford early school achievement test 2 mathematics (SESAT 2)

The Stanford early school achievement test 2 (SESAT 2) mathematics is the kindergarten version of the Stanford achievement test 10th edition series (SAT-10) [19]. The SESAT 2 served as the posttest measure in the current study. The SESAT 2 measures early mathematical skills, including number recognition, counting, and basic operations such as simple addition and subtraction. The internal consistency of SESAT 2 mathematics was reported as .88 in one study [20].

## 2.4. Original research design: RCT

Students were first considered eligible for the ROOTS intervention if they scored 20 or less on the NSB and had a composite ASPENS score that placed them in the strategic or intensive range. The ASPENS and NSB scores of ROOTS-eligible students were then separately converted into standard scores, and the two standard scores were summed to compute an overall composite score. The composite standard scores of ROOTS-eligible students were rank-ordered within each classroom, and the lowest scoring 10 students were then randomly assigned to one of two conditions: a ROOTS group [21] or a no-treatment business-as-usual (BAU) control condition (i.e., random assignment at student-level within each classroom).

## 2.5. Current research design: MCRD

Rank ordering of the composite standard scores in each classroom to identify the 10 lowest-performing students resulted in 106 unique cutoff scores (one from each classroom), enabling the creation of the current synthetic MCRD design [8]. Specifically, from the RCT dataset, a single cutoff based on the assignment variable continuum and the treatment assignment rule were determined, dividing the observations into either the treatment or control condition. Then, an RD treatment (RD-T) group was created by excluding all control group observations in the treated side (e.g., below the cutoff), and an RD control group was created by excluding all treated observations in the control side (e.g., above the cutoff). Following this approach, the MCRD data were constructed from the RCT dataset for each classroom. Specifically, we created an RD treatment (RD-T) group by excluding all control group observations below the cutoff in each classroom while the RD control (RD-C) group in each classroom comprised students whose assignment variable scores fell above the classroom cutoff. In total, 106 unique cutoff scores (one for each classroom) were identified, ranging from -2.17 to 0.24 ( $M=-.75$ ,  $SD=.42$ ). The majority of the classroom cutoff values were below 0, indicating that the ROOTS classrooms were typically below the mean value on the composite screening measure. The RD-T group ( $n=640$ ) and the RD-C group ( $n=1,260$ ) consisted of similar proportions of female and ethnic minority students. The RD-T group included more English language learners and more students receiving special education services than the RD-C group. As expected given the use of a cutoff-based treatment assignment rule, both assignment variable and posttest scores were lower in the RD-T group than in the RD-C group. Table 1 presents counts and demographic information for the two RD groups identified in this study ( $N=1,900$ ).

Table 1. Student characteristics and mean and SD of assignment scores by treatment assignment condition in the MCRD design

Variables	RD-T ( <i>n</i> =640)	RD-C ( <i>n</i> =1,256)
Female <i>n</i> (%)	334 (52.2)	627 (49.92)
Ethnic minority <i>n</i> (%)	70 (11.0)	153 (12.2)
English language learner <i>n</i> (%)	124 (19.4)	127 (10.1)
Student with disabilities <i>n</i> (%)	47 (7.3)	37 (2.95)
Age in years <i>M</i> ( <i>SD</i> )	5.27 (0.44)	5.41 (0.49)
ASPENS composite scores <i>M</i> ( <i>SD</i> )	21.87 (17.50)	74.79 (37.37)
NSB total raw scores <i>M</i> ( <i>SD</i> )	12.22 (3.75)	19.88 (4.80)
Composite standard scores* <i>M</i> ( <i>SD</i> )	-1.52 (0.73)	1.27 (1.40)
Spring SESAT 2 mathematics scores <i>M</i> ( <i>SD</i> )	460.22 (35.43)	494.38 (37.87)

Note. Total MCRD sample across on conditions, *N*=1,900. \*Composite standard scores=composite standard scores created from combining standard ASPENSE and NSB scores

## 2.6. Analytic procedures

Prior to conducting the MCRD analyses, the assignment variable for each classroom was centered at the classroom cutoff, so that all classrooms had a zero-cutoff value. Observations from all classrooms were then merged into a single dataset, resulting in a single centered zero-cutoff (C) across classrooms. This pooled dataset was used for all subsequent MCRD analyses. All analyses for this study were implemented in R [22].

### 2.6.1. Model assumptions

To examine the internal validity of the pooled RD estimate, the primary RD model assumptions for the MCRD design were tested at the centered cutoff (C) [6], [9], [23]. First, we checked whether the distribution of key student characteristics (covariates), such as gender, age, English-language learner status, and special education status was balanced at the cutoff. The examination was conducted by regressing the baseline covariates on the assignment variable and the treatment indicator within an optimal bandwidth that minimizes mean-square error (MSE). Second, to determine if there was manipulation of the assignment scores at the cutoff, we assessed the discontinuity in the probability density function using the McCrary test [24] with the *rddensity* version 2.3 package in R [25].

### 2.6.2. Pooled parametric multiple-cutoff regression discontinuity model

In the MCRD design, the pooled treatment effect represents the average treatment effect across different cut points in the assignment variable continuum [9]. Using the combined (pooled) dataset, we conducted a standard RD analysis at the centered cutoff (C) to estimate this effect. We used a 2-level linear mixed model (i.e., multilevel modeling) [26] with a random intercept and fixed slopes to estimate the treatment effect across classrooms. The multilevel model was used to address the potential dependencies that arise from students (Level 1) nested within classrooms (Level 2). Posttest student SESAT 2 Mathematics scores (Level 1) were regressed on linear, quadratic, and cubic terms of the centered assignment variable, the treatment indicator, and interaction terms between the treatment indicator and centered assignment variable. No predictors were included at the classroom level (Level 2). Predictors not statistically associated with the outcome were removed iteratively, starting with higher-order terms, until the most parsimonious and best-fitting model was identified. We used *F*-statistics and the Akaike information criterion (AIC) value to determine if a more complex model yielded statistically significant improvements. The general form of the parametric pooled MCRD model is specified in (1):

$$Y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - C_j) + \gamma_{20}(x_{ij} - C_j)^2 + \gamma_{30}(x_{ij} - C_j)^3 + \gamma_{40}Z_{ij} + \gamma_{50}(x_{ij} - C_j)Z_{ij} + \gamma_{60}(x_{ij} - C_j)^2Z_{ij} + \gamma_{70}(x_{ij} - C_j)^3Z_{ij} + u_{0j} + r_{ij} \quad (1)$$

Where,  $Y_{ij}$  is the SESAT 2 mathematics score of student  $i$  in classroom  $j$ ;  $x_{ij}$  is the assignment score for student  $i$  in classroom  $j$ ;  $C_j$  is the cutoff for classroom  $j$ ;  $Z_{ij}$  is the dichotomous ROOTS intervention indicator for student  $i$  in classroom  $j$  (0=control, 1=treatment);  $(x_{ij}-C_j)Z_{ij}$  is the interaction between assignment score and the ROOTS intervention indicator for student  $i$  in classroom  $j$ ;  $\gamma_{00}$  is the mean SESAT mathematics score for the control group;  $\gamma_{10}$  is the slope relating the assignment score to the posttest SESAT mathematics score;  $\gamma_{20}$  is the slope relating the quadratic term of the assignment score to the posttest SESAT mathematics score;  $\gamma_{30}$  is the slope relating the cubic term of the assignment score to the posttest SESAT mathematics score;  $\gamma_{40}$  is the difference in the linear slope between the two conditions (0=control, 1=treatment);  $\gamma_{50}$  is the slope for the interaction between the assignment score and the ROOTS intervention indicator;  $\gamma_{60}$  is the slope for the interaction between the quadratic term of the assignment score and the ROOTS intervention indicator;

$\gamma_{70}$  is the slope for the interaction between the cubic term of the assignment score and the ROOTS intervention indicator; and  $r_{ij}$  is the student-level residual term. At the classroom level (Level 2), the intercept ( $u_{0j}$ ) was allowed to vary between classrooms, while all slopes were fixed to be constant across classrooms.

### 2.6.3. Pooled nonparametric multiple-cutoff regression discontinuity model

A nonparametric analysis was conducted using local linear regression (LLR) [13] to supplement the parametric results. First, we identified an MSE-optimal bandwidth ( $h$ ) around the cutoff using the Imbens-Kalyanaraman (IK) procedure [27]. Then, we applied a linear regression with triangular kernel weights within the optimal bandwidth ( $h$ ). The nonparametric pooled RD model is defined in (2):

$$Y_{ij} = \beta_0 + \beta_1(x_{ij} - c_j) + \beta_2 Z_{ij} \\ \text{with triangle kernel weights } K(x) = (1 - |x|)\mathbf{1}\{|x| \leq 1\} \quad (2)$$

Where,  $x_{ij}$  is the assignment score for student  $i$  in classroom  $j$ ;  $c_j$  is the cutoff score for classroom  $j$ ; and  $Z_{ij}$  is the ROOTS intervention indicator for student  $i$  in classroom  $j$  (0=control, 1=treatment). The term  $\beta_2$  represents the magnitude of the treatment effect at the centered cutoff  $C$ . To check the sensitivity of the results to the choice of bandwidth, we repeated the analysis using bandwidths half and twice the optimal size for comparison. Standard errors for the nonparametric estimates were calculated using bootstrapping techniques, with 1,000 repetitions.

### 2.6.4. Multiple-cutoff regression discontinuity heterogeneity model

A 2-level linear mixed model was used to examine potential heterogeneity in the effect of the ROOTS intervention across classrooms. Following the best-fitting parametric pooled model, the base conditional MCRD model was defined as in (3).

$$Y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - c_j) + \gamma_{20}(x_{ij} - c_j)^2 + \gamma_{30} Z_{ij} + \gamma_{31} c_j Z_{ij} + u_{0j} + u_{3j} + r_{ij} \quad (3)$$

Where,  $Y_{ij}$  is the posttest SESAT mathematics score for student  $i$  in classroom  $j$ ;  $x_{ij}$  is the assignment score for student  $i$  in classroom  $j$ ;  $c_j$  is the cutoff score for classroom  $j$ ;  $Z_{ij}$  is the ROOTS intervention indicator for student  $i$  in classroom  $j$  (0=control, 1=treatment);  $\gamma_{00}$  is the classroom mean SESAT mathematics score for the control group;  $\gamma_{10}$  is the slope relating the assignment score to the posttest SESAT mathematics score;  $\gamma_{20}$  is the slope relating the quadratic term of the assignment score to the posttest SESAT mathematics score;  $\gamma_{30}$  represents the difference in the posttest score between the treatment and control groups;  $u_{0j}$  is the random intercept term; and  $r_{ij}$  is the student-level residual term. The treatment effect was modeled to vary across classrooms ( $u_{3j}$ ) to examine the potential heterogeneity of the estimated treatment effect across classrooms. We also included the classroom cutoff value ( $c_j$ ) as a classroom-level predictor to explore whether the classroom assignment cutoff value was associated with amplifying or attenuating the treatment effect ( $\gamma_{31}$ ).

## 3. RESULTS AND DISCUSSION

### 3.1. Results

#### 3.1.1. Assumption tests

Table 2 presents the covariate balance statistics. At the centered cutoff ( $C$ ), there was no statistically significant discontinuity in any of the tested covariates. The McCrary test also revealed statistically non-significant results, suggesting there was no systematic manipulation of the assignment scores around the cutoff.

Table 2. Covariate balance statistics

Covariates	$\Delta$ Difference	SE	$p$	Variance ratio
Female	-.04	.04	.31	1.01
Student age	-.05	.03	.14	.92
Limited English proficiency	.03	.03	.39	1.10
Special education	.01	.02	.67	1.14

Note.  $\Delta$  difference=weighted mean difference in the covariate distribution at the centered cutoff; variance ratio=mean ratio of the variance of a demographic variable in the treatment group to the variance of the variable in the comparison group

#### 3.1.2. Multiple-cutoff regression discontinuity parametric model

Table 3 reports the estimates for the parametric MCRD models. The intraclass correlation coefficient (ICC) indicated that 40% of the variation in student posttest SESAT mathematics scores was attributable to classroom-to-classroom differences. The best-fitting model (Model 1) included a linear and

quadratic term for the assignment variable. The model results indicated a positive relationship between assignment scores and posttest SESAT mathematics scores ( $\gamma_{10}=15.29$ ,  $SE=.89$ ), meaning that an increase in assignment scores predicted a corresponding increase in the posttest SESAT mathematics scores. However, as assignment scores continued to increase, the rate of improvement in posttest SESAT mathematics scores decelerated ( $\gamma_{20}=-.52$ ,  $SE=.17$ ). The RD treatment effect estimate was statistically significant ( $\gamma_{30}=5.30$ ,  $SE=2.32$ ), demonstrating that across the range of cutoff values, students assigned to the treatment condition outperformed those assigned to the control condition on average. Although the addition of a random effect term for the treatment status indicator ( $u_{3j}$ ) did not statistically improve the model (Model 2), the term was retained in order to explore whether treatment effects varied by the cutscore that was used to assign students to the treatment or control condition across classroom units (Model 3).

Table 3. Multilevel SESAT mathematics outcome models ( $N=1,900$ )

Effect	Model 1		Model 2		Model 3	
	Estimate	SE	Estimate	SE	Estimate	SE
Fixed effect						
Intercept, $\gamma_{00}$	468.27***	2.49	468.17***	2.49	478.14***	4.24
Inter-student predictors						
Assignment score, $\gamma_{10}$	15.29***	.89	15.36***	.90	15.92***	.91
Assignment score squared, $\gamma_{20}$	-.52**	.17	-.53**	.17	-.64***	.17
Treatment, $\gamma_{30}$	5.30*	2.32	5.48*	2.37	12.65***	3.60
Inter-classroom predictors						
Cut-point, $\gamma_{01}$					15.30**	5.03
Cut-point, $\gamma_{31}$					8.88*	3.41
Random effect						
Student (level 1), $r_{ij}$	VC	SD	VC	SD	VC	SD
Student (level 1), $r_{ij}$	697.40	21.45	692.05	26.23	692.45	26.31
Mean SESAT score, $u_{0j}$	460.10	26.41	459.43	21.60	419.64	20.49
Treatment effect, $u_{3j}$			23.62	5.35	13.29	3.65

Note: SE=Standard error, VC=Variance components, SD=Standard deviation

\* $p<.05$ , \*\* $p<.01$ , \*\*\* $p<.001$

### 3.1.3. Nonparametric model

Local linear regression was implemented within the MSE-optimal bandwidth (two points) around the cutoff, as determined by the IK procedure. By restricting the analysis to observations within this bandwidth, the local linear analysis used 39.4% of the student observations. Table 4 reveals that the treatment effect within the optimal bandwidth was statistically significant ( $\beta_2=6.63$ ). This result indicated that, on average, students in the treatment condition had higher posttest SESAT mathematics scores than those in the control condition across all cutoff values. The results of the sensitivity analysis also revealed that using the half or double-sized bandwidth led to pooled RD estimates of 8.81 ( $p<.05$ ) and 4.17 points ( $p>.05$ ), respectively, suggesting that the nonparametric impact estimate was sensitive to the bandwidth size. Figure 1 displays the nonparametric local linear regression plot of the assignment and outcome variable relationship, whereby the 6-point discontinuity at the pooled cut point can be seen.

Table 4. Nonparametric pooled RD estimates of SESAT mathematics outcome

Effect	Half bandwidth		Optimal bandwidth		Double bandwidth	
	Estimate	t	Estimate	t	Estimate	t
Intercept, $\beta_0$	465.21 (2.75)	169.08***	466.22 (1.96)	237.67***	467.55 (1.67)	280.40***
Assignment score, $\beta_1$	20.43 (5.26)	3.89***	15.72 (2.03)	7.74***	13.63 (1.05)	13.01***
Treatment, $\beta_2$	8.81 (5.49)	2.30*	6.63 (3.86)	2.24*	4.17 (2.89)	1.63

Note. Nonparametric pooled RD estimates are shown with standard errors in parenthesis. Nonparametric estimates have bootstrapped standard errors (repetition=1,000)

\* $p<.05$ , \*\* $p<.01$ , \*\*\* $p<.001$

### 3.1.4. Multiple-cutoff regression discontinuity heterogeneity model

When the classroom cutoff score was added as a predictor of the variance in treatment effects (see Model 3 and Table 3) in the parametric RD model, results indicated a positive interaction between the cutoff score and the treatment effect parameter ( $\gamma_{31}=8.88$ ,  $SE=3.41$ ). The moderated relationship suggests that higher cutoff scores were associated with greater treatment effects. Notably, the classroom cutpoint was also associated with overall classroom mean performance ( $\gamma_{01}=15.30$ ,  $SE=5.03$ ), indicating that classrooms with higher cutoff values had statistically higher mean posttest scores relative to those with lower cutoff scores.

Figure 2 shows that the local RD treatment effect estimates varied widely depending on the cutoff value used to screen students into treatment conditions, with an approximate range between -57 and +68 points. Classrooms with higher cutoff values generally exhibited larger treatment effects, although the classroom with the lowest cutoff score also displayed a substantial positive effect. It is important to note, however, that the confidence interval for each RD estimate was large due to the small sample size, suggesting that caution should be used when interpreting the individual local RD estimates.

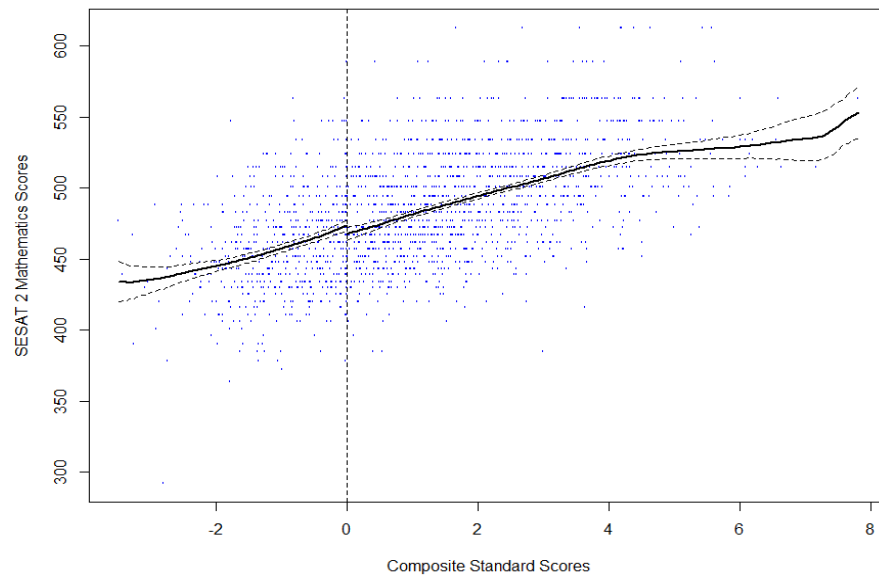


Figure 1. Nonparametric plot of the assignment and outcome variable relationship

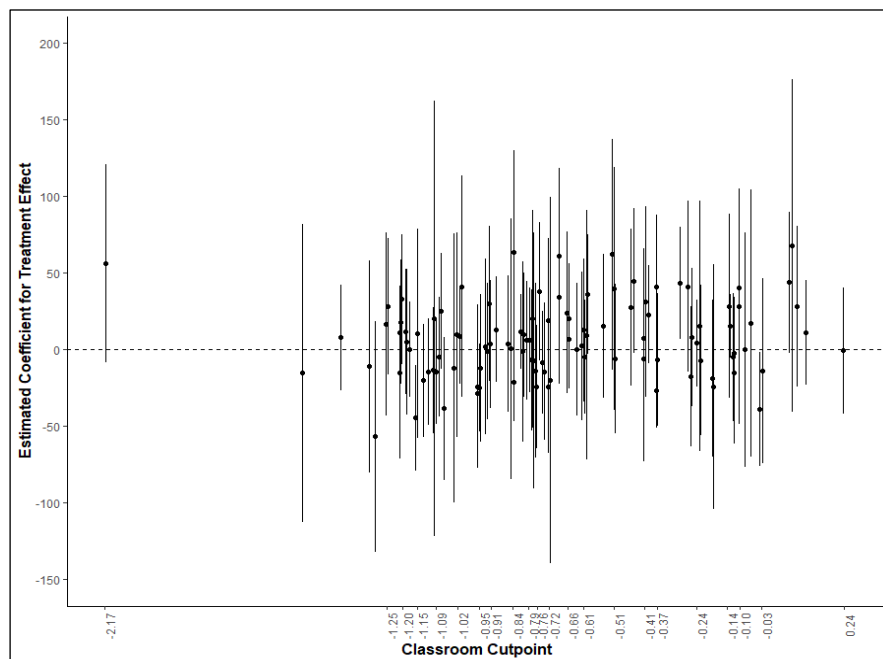


Figure 2. Local RD treatment effect estimates at various classroom cut points with 95% confidence intervals

### 3.2. Discussion

The RD design is often considered as a practical alternative to the RCT in field-based settings as it can provide an unbiased treatment effect estimate when model assumptions are met and the relationship between the assignment and outcome variables is correctly specified. This paper provided an overview of the

MCRD design, an extension of basic RD design used when multiple cutoffs are used program or intervention assignment across different units or sites. Previous analytic approaches for MCRD designs primarily focused on estimating average treatment effects across cutoffs, potentially overlooking site- or unit-specific differences in treatment effects [9]. Conversely, this paper demonstrated a modeling approach that simultaneously estimates both the overall average treatment effect and treatment effect heterogeneity across different intervention units. Findings indicated that students who received the ROOTS intervention had higher mathematics scores on average at post-intervention compared to those who did not. Additionally, the size of the intervention effects varied across classrooms, with greater effects generally observed in classrooms that had higher cutoff scores to screen students for the intervention.

### **3.2.1. Multiple-cutoff regression discontinuity estimation and statistical validity**

The results revealed that the pooled treatment effect estimate varied depending on the estimation method (parametric or nonparametric). The magnitude of the RD estimate was larger when using the nonparametric model (6.48, effect size  $g=.19$ ) [28] compared to the parametric model (5.13, effect size  $g=.14$ ). Although the effect sizes across the methods were relatively similar, the observed differences in the raw treatment effect estimates can be attributed to the methods used in the estimation of each model. The parametric method uses all available observations in the dataset to fit a model, whereas the nonparametric method relies on a subsample available within the optimal bandwidth. In the present demonstration, we did not evaluate the internal validity of the different treatment effect estimates. However, we hypothesize that the nonparametric method is likely to produce a more robust causal estimate than the parametric method, given that it uses only observations near the cutoff. An examination of the standard errors associated with each treatment effect estimate revealed that the nonparametric method produced higher standard errors than the parametric method. This finding is not surprising, given that the nonparametric method uses a smaller number of observations within a small bandwidth to estimate the relationship between the assignment variable and the outcome.

The choice between parametric and nonparametric estimation methods in pooled MCRD analyses will be often context-specific. While the nonparametric model may yield a more robust RD estimate, it generally has lower statistical power compared to the parametric model. Therefore, the nonparametric method is best suited for multi-site program evaluations with relatively large within-unit samples. A large within-site sample will typically be required when the unit-specific treatment effects and RD treatment effect heterogeneity are of primary interest. Additionally, since the nonparametric approaches such as local linear regression utilize the samples within the optimal bandwidth around the cutoff, it is critical to have a sufficient number of observations near the cutoff (i.e., effective sample) to use this method effectively. If a large within-unit sample or effective sample is not available—a common challenge in real-world evaluations—parametric modeling would be a viable alternative, given that it produces an impact estimate with a relatively higher statistical power. We recommend that researchers employ both MCRD estimation methods, along with graphical analysis, to examine the sensitivity of treatment impact and precision estimates provided by each method.

### **3.2.2. Intervention impact heterogeneity in multiple-cutoff regression discontinuity designs**

Results from MCRD model indicated that the ROOTS intervention worked differently depending on the cutoff value chosen to screen students into the treatment condition in each classroom. Specifically, higher assignment score cutoff values were associated with greater intervention effects on average. These findings suggest that the MCRD can be an informative program evaluation design when treatment assignment cutoffs vary by intervention unit or site. By identifying the pooled treatment effect over a range of the assignment variable continuum and examining whether an intervention effect varies at different cut-points, researchers and educators can gain a more comprehensive understanding of program performance and the factors influencing differential outcomes.

The ability to derive a pooled impact estimate across a wider range of the assignment variable continuum notably contrasts with the basic RD design, where causal inference is restricted to a narrow region around a single cutoff score. For educational researchers and program evaluators, the MCRD design may thus be used to probe whether a program is effective or differentially effective across a range of diagnostically important assignment score values. In cases where heterogeneity in the local treatment effects is not observed, a pooled RD estimate serves to provide an accurate summary of the average treatment effect across all cutoffs included in the design. The researcher can then infer the treatment had similar impact across a specified age range or an income, reading score, or preprogram performance intervals, for example. However, if the researcher identifies that the treatment effect varies across different points along the assignment variable continuum, then it becomes possible to investigate whether site characteristics (e.g., unit cutoff value, class size, fidelity of implementation) moderate the size of the treatment impact on the



outcomes of interest. The researcher may then be able to report that program effects varied with respect to more or less malleable program characteristics such as the level of adherence to the program model, the severity of need in the community, or the length of program operations.

The acquisition of more robust insight into the RD treatment effect, including possible heterogeneity, may enable educational program developers and administrators to redesign or modify the delivery of an intervention to better meet the needs of diverse program participants. For example, if an early childhood literacy intervention is shown to have a positive impact for students at moderate risk, but not for those most at-risk, educators could choose to reconsider the intervention's target population, entertain adopting a new program with a better aligned curriculum, or attempt to modify the program to provide a more intense treatment dosage to those most at-risk of negative outcomes. Similarly, in college remediation programs, if a math remediation course benefits students with moderate low scores but not those with the lowest scores, program developers and administrators may choose to reevaluate the program's target population or enhance the program's intensity for those with the lowest scores.

### 3.2.3. Limitations

The current study has several limitations to note. First, we did not evaluate the performance of the MCRD estimate relative to the benchmark RCT causal estimate. That is, the findings of this study do not speak to whether the MCRD design produces treatment effect estimates that are comparable to an RCT benchmark. Therefore, future studies are needed to evaluate the validity and efficiency of MCRD methods in comparison to RCT estimates. Second, the dataset used in this study had a small within-group (classroom) sample size, thereby limiting the reliability of the estimation of local intervention effects at each cutoff. Third, this study considered only one predictor of treatment effect heterogeneity. Given recent research linking group- and classroom-level factors to math development [29], [30], future studies should investigate how various unit-level factors (e.g., instructional strategy, fidelity of implementation, teacher experience) relate to classroom differences in program outcomes.

## 4. CONCLUSION

The motivation for this study was to demonstrate a method for estimating treatment effects when multiple cutoffs are used across different intervention units and to model treatment effect variability within the RD framework for educational evaluation. While further research is needed, the current study suggests that MCRD designs have the potential to identify the pooled intervention effect across cutoffs and uncover the processes and mechanisms through which interventions are differentially effective across individuals and contexts. We recommend that researchers start an MCRD analysis with a traditional pooling analysis to estimate the average treatment effect across cutoffs. Subsequently, they can further explore potential heterogeneity in treatment effects at different cutoff points by examining interaction effects or conducting subgroup analyses based on relevant theory, the research context, and preliminary analytic results (e.g., local RD treatment effect estimates).

## FUNDING INFORMATION

This research was supported by the ROOTS Project (Grant R324A120304), funded by the Institute of Education Sciences, U.S. Department of Education.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
HyeonJin Yoon	✓	✓			✓			✓	✓	✓	✓		✓	
Keith Zvoch	✓	✓				✓			✓	✓	✓	✓	✓	
Keith Smolkowski		✓				✓				✓				✓
Ben Clarke						✓				✓				✓

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

## ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

## DATA AVAILABILITY

Additional data analyses and publications are currently underway. Supporting data is therefore not available.




## REFERENCES

- [1] C. Ong-Dean, C. H. Hofstetter, and B. R. Strick, "Challenges and dilemmas in implementing random assignment in educational research," *American Journal of Evaluation*, vol. 32, no. 1, pp. 29–49, 2011, doi: 10.1177/1098214010376532.
- [2] T. D. Cook, M. Scriven, C. L. S. Coryn, and S. D. H. Evergreen, "Contemporary thinking about causation in evaluation: a dialogue with Tom Cook and Michael Scriven," *American Journal of Evaluation*, vol. 31, no. 1, pp. 105–117, 2010, doi: 10.1177/1098214009354918.
- [3] H. S. Bloom, "Modern regression discontinuity analysis," *Journal of Research on Educational Effectiveness*, vol. 5, no. 1, pp. 43–82, 2012, doi: 10.1080/19345747.2011.578707.
- [4] T. D. Cook, "'Waiting for life to arrive': a history of the regression-discontinuity design in psychology, statistics and economics," *Journal of Econometrics*, vol. 142, no. 2, pp. 636–654, 2008, doi: 10.1016/j.jeconom.2007.05.002.
- [5] J. J. Kirksey and M. A. Gottfried, "The effect of serving 'breakfast after-the-bell' meals on school absenteeism: comparing results from regression discontinuity designs," *Educational Evaluation and Policy Analysis*, vol. 43, no. 2, pp. 305–328, 2021, doi: 10.3102/0162373721991572.
- [6] G. W. Imbens and T. Lemieux, "Regression discontinuity designs: a guide to practice," *Journal of Econometrics*, vol. 142, no. 2, pp. 615–635, Feb. 2008, doi: 10.1016/j.jeconom.2007.05.001.
- [7] P. Z. Schochet, "Statistical power for regression discontinuity design in educational evaluations," *Journal of Educational and Behavioral Statistics*, vol. 34, no. 2, pp. 238–266, Jun. 2009, doi: 10.3102/1076998609332748.
- [8] C. Wing and T. D. Cook, "Strengthening the regression discontinuity design using additional design elements: a within-study comparison," *Journal of Policy Analysis and Management*, vol. 32, no. 4, pp. 853–877, 2013, doi: 10.1002/pam.21721.
- [9] M. D. Cattaneo, L. Keele, R. Titiunik, and G. Vazquez-Bare, "Interpreting regression discontinuity designs with multiple cutoffs," *Journal of Politics*, vol. 78, no. 4, pp. 1229–1248, 2016, doi: 10.1086/686802.
- [10] Y. K. Önder and M. Shamsuddin, "Heterogeneous treatment under regression discontinuity design: application to female high school enrolment," *Oxford Bulletin of Economics and Statistics*, vol. 81, no. 4, pp. 744–767, 2019, doi: 10.1111/obes.12292.
- [11] W. van der Klaauw, "Breaking the link between poverty and low student achievement: an evaluation of Title I," *Journal of Econometrics*, vol. 142, no. 2, pp. 731–756, 2008, doi: 10.1016/j.jeconom.2007.05.007.
- [12] K. Y. Chay, P. J. McEwan, and M. Urquiola, "The central role of noise in evaluating interventions that use test scores to rank schools," *American Economic Review*, vol. 95, no. 4, pp. 1237–1258, 2005, doi: 10.1257/0002828054825529.
- [13] M. D. Cattaneo and R. Titiunik, "Regression discontinuity designs," *Annual Review of Economics*, vol. 14, no. 1, pp. 821–851, 2022, doi: 10.4135/9781529682717.n7.
- [14] R. E. Slavin, "What works? Issues in synthesizing educational program evaluations," *Educational Researcher*, vol. 37, no. 1, pp. 5–14, 2008.
- [15] S. M. Loftus-Rattan, M. Wrightington, J. Furey, and J. Case, "Multi-tiered system of supports: an ecological approach to school psychology service delivery," *Teaching of Psychology*, vol. 50, no. 1, pp. 77–85, 2023, doi: 10.1177/00986283211024262.
- [16] B. Clarke, R. Gersten, J. Dimino, and E. Rolffhus, *Assessing student proficiency in early number sense (aspens) administrator's handbook*. Colorado: Cambium Learning Group, Sopris, 2012.
- [17] N. C. Jordan, J. Glutting, and C. Ramineni, "A number sense assessment tool for identifying children at risk for mathematical difficulties," in *Mathematical Difficulties: Psychology and Intervention*, A. Dowker and G. D. Phye, Eds., San Diego, CA: Academic Press, 2008, pp. 45–58, doi: 10.1016/B978-012373629-1.50005-8.
- [18] N. C. Jordan, J. Glutting, and C. Ramineni, "The importance of number sense to mathematics achievement in first and third grades," *Learning and Individual Differences*, vol. 20, no. 2, pp. 82–88, 2010, doi: 10.1016/j.lindif.2009.07.004.
- [19] Harcourt Educational Measurement, *Stanford achievement test series*. New York: Harcourt Educational Measurement, 2003.
- [20] C. T. Doabler *et al.*, "Testing the efficacy of a Tier 2 mathematics intervention: a conceptual replication study," *Exceptional Children*, vol. 83, no. 1, pp. 92–110, 2016, doi: 10.1177/0014402916660084.
- [21] C. T. Doabler, B. Clarke, D. Kosty, E. Kurtz-Nelson, H. Fien, K. Smolkowski, and S. K. Baker, "Examining the Impact of Group Size on the Treatment Intensity of a Tier 2 Mathematics Intervention Within a Systematic Framework of Replication," *Journal of Learning Disabilities*, vol. 52, no. 2, 2018, doi: 10.1177/0022219418789376.
- [22] F. M. Giorgi, C. Ceraolo, and D. Mercatelli, "The R language: an engine for bioinformatics and data science," *Life*, vol. 12, no. 5, p. 648, 2022, doi: 10.3390/life12050648.
- [23] D. S. Lee and T. Lemieux, "Regression discontinuity designs in economics," *Journal of Economic Literature*, vol. 48, no. 2, pp. 281–355, 2010, doi: 10.1257/jel.48.2.281.




- [24] J. McCrary, "Manipulation of the running variable in the regression discontinuity design: a density test," *Journal of Econometrics*, vol. 142, pp. 698–714, 2008.
- [25] M. D. Cattaneo, M. Jansson, and X. Ma, "Manipulation testing based on density discontinuity," *The Stata Journal*, vol. 18, no. 1, pp. 234–261, 2018, doi: 10.1177/1536867X1801800115.
- [26] D. A. Luke, *Multilevel modeling*, 2nd ed. Thousand Oaks, CA: SAGE Publications, 2019, doi: 10.1007/978-3-319-57413-4\_18.
- [27] G. Imbens and K. Kalyanaraman, "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of Economic Studies*, vol. 79, no. 3, pp. 933–959, Jul. 2012, doi: 10.1093/restud/rdr043.
- [28] J. M. Taylor and S. Alanazi, "Cohen's and hedges' g," *Journal of Nursing Education*, vol. 62, no. 5, pp. 316–317, May 2023, doi: 10.3928/01484834-20230415-02.
- [29] L. Cragg and C. Gilmore, "Skills underlying mathematics: the role of executive function in the development of mathematics proficiency," *Trends in Neuroscience and Education*, vol. 3, no. 2, pp. 63–68, 2014, doi: 10.1016/j.tine.2013.12.001.
- [30] A. H. Hindman, L. E. Skibbe, A. Miller, and M. Zimmerman, "Ecological contexts and early learning: contributions of child, family, and classroom factors during head start, to literacy and mathematics growth through first grade," *Early Childhood Research Quarterly*, vol. 25, no. 2, pp. 235–250, 2010, doi: 10.1016/j.ecresq.2009.11.003.

## BIOGRAPHIES OF AUTHORS






**HyeonJin Yoon**    is a research assistant professor at the Nebraska Center for Research on Children, Youth, Families, and Schools, at the University of Nebraska-Lincoln, USA. Her primary research focuses on designing and evaluating educational and behavioral intervention programs for children, students, educators, and families, using rigorous experimental and quasi-experimental research designs. Additionally, she examines innovative methodological approaches to research designs applicable in natural, real-world settings. Dr. Yoon has extensive experience in educational measurement and assessment, particularly in the development and validation of reading assessments for K-12 students and post-secondary learners. She can be contacted at email: hyoon5@unl.edu.






**Keith Zvoch**    is a professor in the Department of Education Studies at the University of Oregon (UO), USA. His research interests focus on causal inference in applied field settings, methodological approaches for measuring and addressing breakdowns in intervention implementation, and longitudinal data analysis. Dr. Zvoch has published in evaluation, education, and child development journals and has served as an associate editor for the American Journal of Evaluation. Dr. Zvoch has led and co-led three large-scale grants funded by the U.S. Department of Education. He can be contacted at email: kzvoch@uoregon.edu.



**Keith Smolkowski**    is a senior scientist at the Oregon Research Institute, USA. He has extensive expertise in designing and analyzing complex efficacy and effectiveness trials across the social science, education, and public health research domains. His research interests focus on various areas including mathematics and reading instruction, interventions for child social behavior, substance use, adolescent problem behavior, and teacher and student assessment. His research contributions have led to over 100 peer-reviewed articles and book chapters. Dr. Smolkowski has led and co-led over 50 research grants funded by U.S. federal agencies, including the U.S. Department of Education, National Science Foundation, and the National Institutes of Health. He can be contacted at email: keiths@ori.org.



**Ben Clarke**    is a professor in the School Psychology Program at the University of Oregon (UO), USA. His research centers on the mathematical development of young children who are with or at-risk for mathematics learning difficulties. With over 20 U.S. federally funded research grants, Dr. Clarke leads intervention programs and develops assessment materials designed to prevent mathematics difficulties. He also applies his research to teaching and mentoring students, guiding them to integrate an understanding of mathematics development into school-based service delivery systems. Dr. Clarke has authored numerous articles and book chapters on mathematics instruction and assessment. Prior to his academic career, he served as a school psychologist, leading efforts to implement tiered instructional models in reading and mathematics. He can be contacted at email: clarkeb@uoregon.edu.