MULTIPLE-CUTOFF REGRESSION DISCONTINUITY DESIGNS IN PROGRAM EVALUATION: A COMPARISON OF TWO ESTIMATION METHODS

by

HYEONJIN YOON

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2018

DISSERTATION APPROVAL PAGE

Student: HyeonJin Yoon

Title: Multiple-cutoff Regression Discontinuity Designs in Program Evaluation: A Comparison of Two Estimation Methods

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Keith Zvoch	Chairperson
Gina Biancarosa	Core Member
Keith Smolkowski	Core Member
John R. Seeley	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2018

DISSERTATION ABSTRACT

HyeonJin Yoon

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

September 2018

Title: Multiple-cutoff Regression Discontinuity Designs in Program Evaluation: A Comparison of Two Estimation Methods

In basic regression discontinuity (RD) designs, causal inference is limited to the local area near a single cutoff. To strengthen the generality of the RD treatment estimate, a design with multiple cutoffs along the assignment variable continuum can be applied. 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 study is to (a) demonstrate the application of two treatment effect estimation methods (i.e., a conventional pooling method and a multilevel pooling method) for the multiple-cutoff RD (MCRD) designs using Tier 2 kindergarten math intervention data (ROOTS), (b) examine the extent to which the two methods yield unbiased and precise estimates comparable to those from the randomized controlled trial (RCT) design, and (c) investigate the moderating role of a classroom characteristic (i.e., classroom cut-point) on the size of the ROOTS intervention effect.

Math intervention data were collected from 2012 to 2015 to evaluate the impact of a small-group (Tier 2) kindergarten mathematics intervention. The analytic sample included 1,900 kindergarten students from the four school districts in Oregon and from the two districts in Boston, Massachusetts. The intervention effect was estimated using a conventional pooling method and a multilevel pooling method. The bias and power of the resulting MCRD estimates were compared with an RCT benchmark. In addition, treatment effect variability was predicted by the cut-point used to screen treated students in each classroom.

Results showed that treatment students scored higher on the posttest outcome than control students at the centered cutoff. All of the MCRD methods produced unbiased treatment effect estimates comparable to a benchmark RCT estimate; however, the power in the MCRD design was lower than in the RCT, regardless of the estimation method. The cut-point used to screen students into the treatment condition moderated the treatment effect, with a greater treatment effect observed in the classrooms with a larger cutoff value. Implications for program evaluation design theory and practice are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: HyeonJin Yoon

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Hanyang University, Seoul, Republic of Korea
Hankuk University of Foreign Studies, Seoul, Republic of Korea

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2018, University of Oregon
Specialization: Quantitative Research Methods
Master of Arts, Education, 2012, Hanyang University
Bachelor of Arts, English and American Literature, 2009, Hankuk University of Foreign Studies

AREAS OF SPECIAL INTEREST:

Quantitative Research Methods
Program Evaluation
Measurement and Assessment
Individual differences in reading compression difficulties

PROFESSIONAL EXPERIENCE:

Research Assistant, Center on Teaching and Learning, 2018-Present

Graduate Employee, Center on Teaching and Learning, 2014-2018

English Teacher, Gyomun Middle School, 2012-2013

GRANTS, AWARDS, AND HONORS:

College of Education Travel Award, University of Oregon, 2017

1st Place Poster Session Winner, University of Oregon, 2017

EMPL Travel Grant Award, University of Oregon, 2014-2018

Ken A. Erickson Memorial Scholarship, University of Oregon, 2015

BRICS Research Grant, Hankuk University of Foreign Studies, 2006

PUBLICATIONS:

Kucheria, P., Sohlberg., M. M., & Yoon, H., Fickas, S., & Prideaux, J. (in press). Read, Understand, Learn, & Excel (RULE): Development and feasibility of a reading comprehension measure for postsecondary learners. *American Journal of Speech-Language Pathology*.

Biancarosa, G., Kennedy, P., Carlson, S. E., Yoon, H., Seipel, B., Liu B., & Davison, M. (2018). Constructing subscores that add to validity: A case study of identifying students at-risk. *Educational and Psychological Measurement*. doi: 10.1177/0013164418763255

Yoon, H., & Jang, H. (2012). Effects of transformational leadership from principals on elementary school teachers' organizational citizenship behavior: Mediation effects of faculty trust in colleagues. *Korean Journal of Educational Administration*, *30*(4), 326-348.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support of many people. I wish to express sincere appreciation to my advisor Dr. Keith Zvoch for his excellent assistance and guidance in the preparation of this manuscript. My thanks also go to Dr. Gina Biancarosa, who mentored me in applied research settings and provided me with tremendous opportunities, and to my committee members Drs. John Seeley and Keith Smolkowski for their invaluable feedback on my dissertation. I would also like to thank my mentors and colleagues at the Center on Teaching and Learning, particularly Drs. Sarah Carlson and Lina Shanley, who always believed in and encouraged me to keep going. I wish to thank Dr. Ben Clarke for supporting me to use the ROOTS data to conduct my dissertation study.

In addition, I would like to express my heartfelt gratitude to the Eugene Korean catholic community, particularly Helen and Sean, who have taken care of me like a family member since I moved to Eugene. I would not have come this far without the constant support, encouragement and inspiration of my best friends Jane, Sunhi, and HyeJin; thank you, ladies. I thank my parents Won-Hee and Byung-Gu, who taught me persistence, courage, and love. I am blessed to have them in my life. Finally, I thank God for being and walking with me all the way.

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

For my parents Won-Hee and Byung-Gu, who always loved and believed in me.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The regression discontinuity (RD) design has been widely recommended as an alternative means to facilitate causal inference (Bloom, 2012). Recently, the Institute of Education Sciences (IES) affirmed the use of RD as a method to evaluate educational intervention programs, claiming that if statistical modeling assumptions are met and the design is properly implemented and analyzed, RD yields an unbiased treatment effect estimate (Jacob, Zhu, Sommers, & Bloom, 2012). In RD designs, individuals or units are assigned to either the treatment or control condition based on a cutoff score on a preprogram measure. The assignment variable can be any measure taken prior to treatment, including the pretest scores of the outcome variable. The assignment variable does not have to be highly correlated with the outcome variable, but the design has more statistical power if it does. In many applications, the assignment variable assesses the participants' need for the treatment or intervention. For example, for a college math remedial program, students' prior math test scores, such as the SAT mathematics scores, are utilized.

Figures 1 and 2 present the scatter plots of assignment variable scores against posttest scores used in the RD designs. Both scatter plots indicate a cutoff set at 50; subjects who score above the cutoff are assigned to the treatment condition, and those who score below the cutoff are assigned to the control condition. Figure 1 depicts a situation where a treatment has no effect. When the treatment is not effective, no discontinuity in the regression relationship between the assignment variable and posttest scores at the cutoff is observed. In contrast, when the treatment is effective, as shown in

Figure 2, a discontinuity in the regression relationship between the assignment variable and posttest scores occurs at the cutoff as the effect estimate is added to the posttest scores for all treatment participants. The difference in the intercept between the treatment and control regression segments at the cutoff indicates the magnitude of the treatment effect.

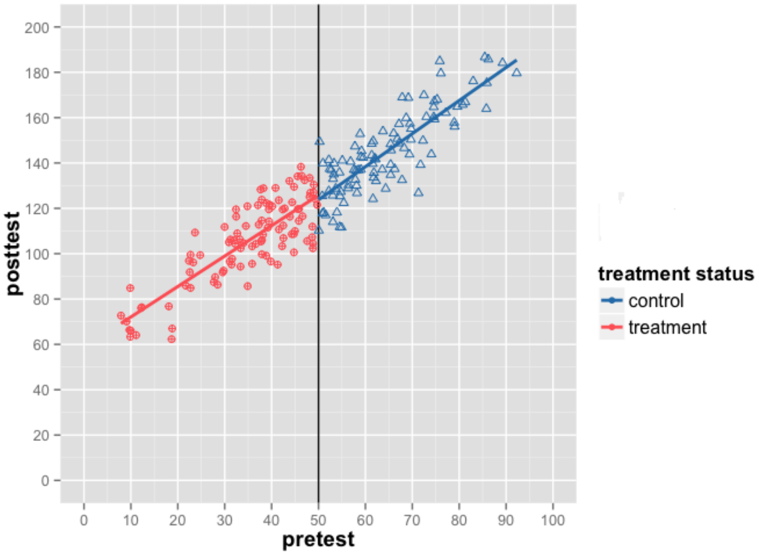


Figure 1. RD with no treatment effects.

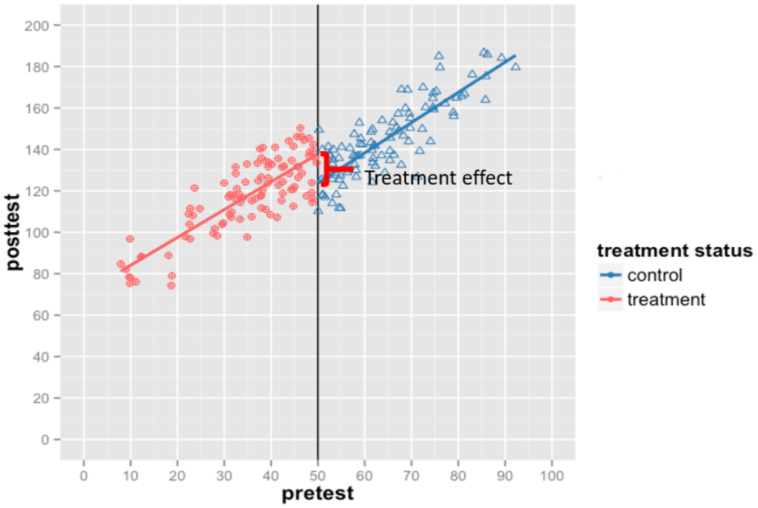


Figure 2. RD with treatment effects.

RD offers three major advantages in a program evaluation context. First, it enables program administrators to target those who are most in need of treatment. As noted earlier a cut-score on the assignment variable is identified and those above or below the cutoff receive the treatment.

Second, RD enables the estimation of an unbiased causal estimate because the selection mechanism is completely known. Specifically, in RD, the use of a cutoff score to assign participants to the treatment and control conditions results in two nonequivalent groups: those with assignment scores below the cutoff and those with assignment scores above the cutoff. However, the source of the nonequivalence (i.e., the assignment score) is statistically modeled. Modeling the assignment scores adjusts for the group differences, thereby producing an unbiased causal estimate at the cut-score (Jacob et al., 2012). In addition, a small area near the cutoff, treatment assignment either to the treatment or control condition can be considered random due to a measurement error in the assignment variable. The local randomization also provides additional conceptual support for unbiased treatment effect estimation at the cutoff.

For these reasons, several authors have recommended RD as a strong alternative to RCT, especially when the random assignment of individuals most in need of treatment is considered neither ethical nor feasible (Bloom, 2012; Schochet et al., 2010; Smith, 2014). Accordingly, RD has been applied to evaluate educational intervention programs, including Reading First (Gamse, Jacob, Horst, Boulay, & Unlu, 2008) and Head Start (Ludwig & Miller, 2007; Wong, Cook, Barnett, & Jung, 2008), as well as post-secondary remedial education programs (Calcagno & Long, 2008; Jacob & Lefgren, 2004; Moss, Yeaton, & LIoyd, 2014).

Despite the advantages offered by RD, the basic RD design has several methodological limitations, including the following: (a) lower statistical power than a comparable RCT, (b) strong reliance on correct modeling of the assignment variable-outcome relationship, and (c) the limited generality of causal inference (Tang, Cook, Kisbu-Sakarya, Hock, & Chiang, 2017; Wing & Cook, 2013). To date, some efforts have been made to mitigate these challenges through supplemental analytic methods or by employing RD design variations. For example, to correctly specify the functional form in RD, researchers have recommended using nonparametric methods in addition to parametric methods (Bloom, 2012; Hahn, Todd, & Van der Klaauw, 2002; Imbens & Lemieux, 2008; Lee & Lemieux, 2010; Sack & Ylvisaker, 1978). Nonparametric methods do not require the pre-specification of the functional form of the estimated relationship between the assignment variable and the outcome and enable the functional form to be determined by the data. Similarly, adding an untreated pretest function or another untreated comparison group has been suggested to improve the precision of RD because these approaches increase the sample size of the study and partial the correlation between the treatment status and the assignment variable (Tang et al., 2017; Wing & Cook, 2013).

An important aspect of these efforts has been the increased attention paid to improving the generality of RD estimates beyond the cutoff. One approach is to extend the area of causal inference in RD designs by using multiple cutoffs across the assignment variable continuum (Black, Galdo, & Smith, 2007; Cattaneo, Titiunik, Vazquez-Bare, & Keele, 2016; Gamse et al., 2008), which is the focus of this study. Multiple-cutoff RD (MCRD) designs arise from variations in the cutoff used to assign individuals or units to conditions (e.g., the school, district, and state). For example, colleges may have different GPA cutoffs to select merit-based scholarship recipients. Similarly, school districts may use different state assessment score cutoffs as the eligibility criteria for summer literacy programs. The availability of multiple cutoffs spread over an assignment variable continuum extends the area of causal inference in RD designs beyond the vicinity of a single cutoff. Thus, in MCRD designs, researchers can estimate an average RD treatment effect by pooling treatment effect estimates across cutoffs, or they can estimate the local RD treatment effects at each cutoff separately.

Although MCRD is a promising approach for generalizing the RD treatment effect estimate, the current literature lacks a thorough examination of the estimation methods used in the application of this design. A common approach to identify treatment effects in the MCRD designs is to estimate a pooled RD treatment effect across cutoffs (Cattaneo et al., 2016). However, the statistical conclusion validity of this approach and the resulting implications for causal inference have not been well established. Similarly, there is also a relative dearth of research on the alternative MCRD method (i.e., multilevel pooling RD) that that allows for the estimation of and modeling of heterogeneity in treatment effects across cutoffs. As a result of the unanswered questions surrounding the MCRD design and estimation, the primary goals of this study are as follows:

1. Demonstrate two treatment effect estimation methods for MCRD designs
2. Evaluate the two estimation methods with respect to the validity, efficiency, and generalization of the causal inference
3. Estimate and model treatment effect heterogeneity in MCRD designs

The findings of this study are expected to contribute to the growing body of literature on RD design variations intended to improve the generality of causal effect estimates. The results of this study may also help program providers and policy makers identify a valid, efficient, and informative program evaluation model when treatment assignment cutoffs vary by sites or units in practice.

In the following, I provide an overview of the theoretical framework and previous studies on RD. I specifically discuss the research literature pertaining to early mathematics intervention programs, which constitutes the applied context for the current study. In addition, I review the previous research on MCRD designs and treatment effect heterogeneity. Then, I discuss the limitations in the current literature on MCRD and early mathematics intervention programs before presenting the research questions for this study.

CHAPTER II

LITERATURE REVIEW

Regression Discontinuity (RD) Designs

In the RD design, subjects on one side of a cutoff score are assigned to the treatment group and those on the other side are assigned to the control group (Bloom, 2012; Shadish, Cook, & Campbell, 2002). For example, applying the RD framework to individuals selected for a Medicaid benefit (i.e., a health coverage program jointly funded by the federal and state governments), the program assigns those who score below an income cutoff in a base period to the treatment group and those who score above the cutoff to the control group. The RD design approach involves a comparison between the two groups with respect to the assignment-outcome variable relationship (e.g., income-medical expenditure). A discontinuity at the cut in the assignment-outcome regression line yields a local causal inference of the treatment effect on the outcome measure (i.e., the effect of Medicaid funds on medical expenditure).

In addition to causal inference, another compelling feature of RD design is associated with the ethical advantage it may offer in real-world, practical settings. Although the RCT provides the most straightforward means of identifying a treatment effect (Odom et al., 2005; What Works Clearinghouse, 2017), it places a burden on program administrators with respect to the randomization of subjects who are most in need of treatment. For example, when implementing a college remedial math program, program administrators may find it unethical to randomize students who score the lowest into either the treatment or control condition. In such scenarios, the use of RD enables program administrators to offer the program to those most in need. In the following paragraphs, I provide a detailed description of the theoretical mechanism facilitating causal inference from RD designs.

Theoretical Framework of RD Designs

Thistlethwaite and Campbell (1960) first proposed and demonstrated RD as an approach for drawing causal inferences in an educational research context where participants were not randomly assigned to experimental and control groups. In this study, students received scholarships based on their score on a scholarship qualifying test, and the authors estimated the effect of student scholarships on career aspirations at the cutoff score for the award. Until 1970s, a major focus of the research was on the theorization of causal inference in RD designs along with the limitations (Campbell, 1969; Goldberger, 1972; Riecken et al., 1974). The identification of the limitations in RD designs led to significant methodological advances, particularly in econometrics (e.g., Angrist, Imbens, & Rubin, 1999; Imbens & Angrist, 1994). Although there were a couple of decades in which RD designs received less attention, since its initial introduction in the 1960s, RD designs have been well-established in both theory and methodology by the concerted efforts of researchers across different disciplines, including education, psychology, econometrics, and mathematics (Cook, 2008).

The primary rationale for RD as a methodology of estimating causal inference is that a completely known assignment rule is used to address the selection bias that is otherwise inherent in the design (Campbell & Stanley, 1963). Goldberger (1972) proved that incorporating the assignment variable in the estimation of the treatment effect could remove the confounding due to selection bias. In other words, he corroborated the notion that the treatment status indicator loses any “explanatory power with respect to an outcome (*y*)" (Goldberger, 1972, p. 16) when including the assignment variable (*x*) in the regression because the assignment variable (*x*) completely determines treatment status (*z*). This means that when incorporating the assignment variable (*x*) in the regression model, the partial correlation between the treatment status (*z*) and the outcome becomes zero. Hence, when the RD model includes the assignment variable, the selection procedure does not result in a spurious effect, thereby yielding an unbiased treatment effect estimate.

Similarly, Campbell and Stanley (1963) suggested that in RD designs, the chance of individuals around the cutoff being assigned either into treatment or control condition is random due to measurement error. Given the assumption of "local randomization," individuals immediately above and below the cutoff are assumed to be identical, except in terms of the treatment assignment status. Therefore, any difference in mean outcomes between these two groups near the cutoff should only be attributed to the treatment (Jacob et al., 2012), which supports unbiased causal inference in RD designs (Lee & Lemieux, 2010).

Methodological Limitations in RD Designs

Once the theoretical framework of RD had been established, researchers started identifying the methodological limitations, including (a) a strong reliance on correct specification of the regression function, (b) low statistical power, and (c) a lack of generality of causal inference (Bloom, 2012; Wing & Cook, 2013; Tang et al., 2017). In response, the researchers made efforts to mitigate these challenges by using supplemental analytic methods (Angrist & Pischke, 2009; Hang et al., 2001; Imbens & Lemieux, 2008; Sack & Ylvisaker, 1978; Schochet, 2009) or by experimenting with alternative RD designs, such as the comparative RD (CRD) design (Wing & Cook, 2013, Tang et al., 2017) and the hybrid RD design (Shadish et al., 2002; Trochim, 1984).

Strong reliance on the correct specification of the regression function. In RD, the validity of the causal estimate relies heavily on correct specification of the regression function linking the assignment and the outcome variable. For example, if researchers modeled a linear function when the true function for the hypothesized relationship is not linear (e.g., curvilinear), an artifactual discontinuity at the cutoff could be observed (Lee & Lemieux, 2010).

In light of this strong reliance on identifying the correct functional form, researchers have suggested using supplemental analytic methods in addition to parametric methods. Sack and Ylvisaker (1978) introduced local nonparametric methods in analyzing RD data. In this approach, a separate linear slope is estimated for ranges of local values of the assignment variable, and greater weights are assigned to observations near the cutoff. Since their introduction into RD analysis, nonparametric methods have widely been used (Angrist & Pischke, 2009; Bloom, 2012; Hahn et al., 2001; Imbens & Lemieux, 2008; Lee & Lemieux, 2010). In nonparametric methods, a particular functional relationship between the predictor and the outcome variable is not predetermined as in parametric methods but is rather constructed according to information from the data. Therefore, nonparametric methods allow for flexibility in modeling functional form across the assignment variable continuum (Bloom, 2012; Whitley & Ball, 2002). In many RD analyses, nonparametric methods have been used to supplement the treatment effect estimates from parametric methods, thereby allowing researchers to assess the potential misspecification of parametric functional forms (Moss et al., 2014; Shadish, Galindo, Wong, Steiner, & Cook, 2011; Wing & Cook, 2013; Wong et al., 2008). The expectation is that consistent results found across parametric and nonparametric methods strengthen the validity of the treatment effect estimate (Lee & Lemieux, 2010).

Low statistical power. Low statistical power due to the correlation between the treatment status indicator and the assignment variable is another area that researchers have attempted to address in RD designs (Goldberger, 1972; Schochet, 2009; Tang & Cook, 2014). Goldberger (1972) demonstrated that RD has a lower statistical power than an RCT, as it requires a sample size between 2.75 and 4 times greater than that of a comparable RCT to detect the same treatment effect. Lower statistical power would be a secondary concern in RD studies with large-scale datasets. However, it would be a major issue in small-scale RD studies, in which individual researchers or small school districts collect their own data (Wing & Cook, 2013). Therefore, to improve power, researchers have suggested adding covariates, pretest scores, or untreated comparison group observations to the model (Calonico, Cattaneo, Farrell, & Titiunik, 2016; Schochet, 2009; Tang & Cook, 2014; Wing & Cook, 2013).

Specifically, covariates are often included in the basic RD design to increase precision, as is common in the analysis of randomized experiments (Calonico et al., 2016). The expectation is that the covariates in the RD regression will reduce unexplained outcome variance. Recent work by Calconico and colleagues (2016) proved that including covariates in RD estimation can lead to an improvement in precision, without substantially altering the RD estimate.

The addition of pretest scores or untreated comparison group observations in the basic RD design also improves efficiency by increasing the sample size for the analysis and reducing the correlation between the treatment status variable and the assignment (Schochet, 2009; Wing & Cook, 2013). Wing and Cook (2013) revealed that the inclusion of pretest scores produced RD estimates that were more precise than those from the basic RD design, and it yielded standard errors close to those of a comparable RCT.

Lack of generality of causal inference. As noted earlier, one key limitation of basic RD is the limited generality of the treatment effect (Tang et al., 2017; Wing & Cook, 2013; Zvoch, Yoon, & Cook, 2016). In the basic RD design, knowing the shape of the treatment regression function in the untreated part or of the control group function in the treated part of the assignment variable continuum is not possible. As a result, causal inference in basic RD is limited to the small area surrounding the cutoff where local randomization occurs.

To strengthen the generality of the RD treatment effect, researchers have recently experimented with several RD design variants, including the use of comparative design elements (Tang et al., 2017; Wing & Cook, 2013; Wing & Bello-Gomez, 2018), the incorporation of a randomization interval in the basic RD design (Black et al., 2007; Cappelleri & Trochim, 1994, 1995; Moss et al., 2014; Trochim & Cappelleri, 1992; Zvoch et al., 2016), and the use of multiple cutoffs (Black et al., 2007; Cattaneo et al., 2016; Gamse et al., 2008).

Comparative RD Designs

CRD is a design in which observations on pretest outcomes (CRD-pre) or comparison groups (CRD-CG) are added to the basic RD design (Tang et al., 2017; Wing & Cook, 2013). As can be seen in Figure 3, CRD comprises the following three regression segments: (a) an RD treated segment (light-pink line), (b) an RD control segment (blue line), and (c) a comparative segment (green line). The comparison observations are from pretest scores or the *untreated* comparison group (e.g., posttest scores from the previous cohort, which was *not* exposed to treatment). The key assumption in the CRD is that the comparative regression function is parallel with the RD control function, and the gap between the two functions is invariant across (above the cutoff in this example) (Tang et al., 2017; Wing & Bello-Gomez, 2018). If the parallel assumption is met, the RD control regression function is extrapolated toward the treated part of the assignment variable continuum (dotted green line in Figure 3), under the assumption that the fixed difference between the RD control function and the comparative function in the control part of the assignment variable continuum (above the cutoff in Figure 3) continues to exist in the treated part (Wing & Bello-Gomez, 2018). The hypothetical counterfactual function created in the treated part of the assignment variable continuum allows the estimation of the average treatment effect on the treated (ATT), thereby supporting the extrapolation of the causal inference of treatment effects beyond the cutoff.

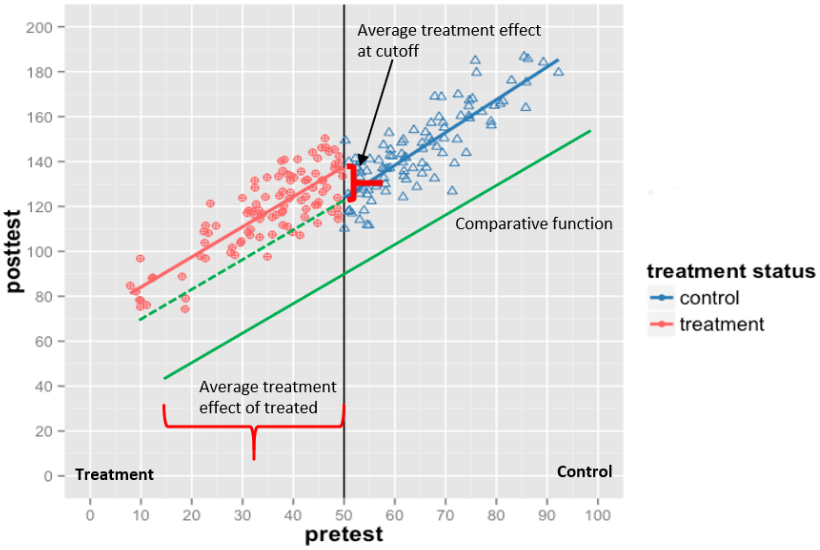


Figure 3. Average treatment effect at cutoff and average treatment effect on the treated using comparative untreated observations.

In a CRD application, Tang et al. (2017) found that adding pretest scores and comparative group observations to the basic RD produced unbiased causal estimates above the cutoff as efficient as RCT and more efficient than the basic RD design did. Specifically, in Tang et al.'s study (2017), as the regression functional form of the comparative untreated observations was parallel with that from actual untreated observations in the basic RD, the treatment effect estimates yielded at the cutoff could be correctly extrapolated above the cutoff. The inclusion of additional comparative cases also improved the precision of the resulting treatment effect estimates. The addition of pretest scores increased the sample size and also reduced the correlation between the treatment status variable and the assignment variable, which increased precision by reducing the standard error of the treatment estimate (Schochet, 2009; Wing & Cook, 2013). Given the advantages that CRD offers with respect to increasing the generality of causal inference and improving precision, researchers have strongly recommended that program evaluators and policy makers use CRD instead of the basic RD whenever possible (Tang & Cook, 2014; Tang et al., 2017).

Hybrid RD Designs

Hybrid RD (HRD; Shadish et al., 2002; Trochim, 1984), another variant of basic RD that was developed to increase the generality of RD estimates, combines the basic RD design with an RCT. To implement an HRD, a randomized interval along the assignment variable continuum is first identified. Second, participants falling into that interval are randomly assigned either to the treatment or control group. Third, participants below or above the cutoffs are assigned to either the RD treatment or the RD control condition (see Figure 4).

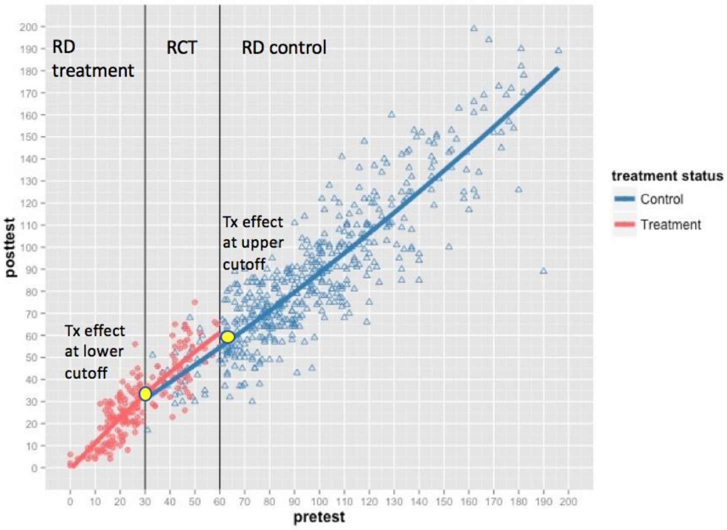


Figure 4. Treatment effect at the upper cutoff in an HRD.

In the HRD design modeled in Figure 4, for example, all participants scoring below the lower cutoff are assigned to the RD treatment condition, while those scoring above the upper cutoff are assigned to the RD control condition (Shadish et al., 2002). Participants whose assignment score falls between cutoffs (randomization interval) are randomly assigned to either the RCT treatment or the RCT control group. In an HRD framework, RD estimates at the lower and upper cutoffs, as well as the average treatment effect within the randomization interval, can be estimated if the regression functions relating the assignment variables to the outcomes for the treatment and control groups are parallel across the randomized interval. If the regression functions are not parallel (i.e., an interaction between an assignment variable and outcome variable is observed), differential treatment effects can be estimated depending on the value of the assignment variable (Zvoch et al., 2016).

HRD improves causal generalization by embedding a randomization interval into the basic RD design. The availability of treatment and control cases in the randomization interval has the advantage of allowing the estimation of actual assignment–outcome regression functions for both treatment and control cases over a common part of the assignment variable continuum. The estimation of the average treatment effects (ATE) within the randomization interval then supports the causal generalization within the randomization interval. In addition, inclusion of the randomization interval also increases precision by reducing the correlation between treatment status and the assignment variable.

Both CRD and HRD improve the fundamental limitations of the basic RD: (a) limited causal generalization and (b) lower statistical power. However, a key difference between HRD and CRD is that the causal generalization in HRD is still limited within the randomization interval, whereas in CRD the causal generalization can be made along the entire assignment variable continuum. The advantage that HRD provides over CRD is that it relaxes a stringent statistical assumption that is required in the latter approach. In a CRD design, the regression functional form for the comparative controls has to be parallel to the regression function for the observed control cases. In an HRD design, again, the availability of actual treatment and control cases within the randomization interval enables the estimation of the treatment effect within the interval. Therefore, HRD does not necessitate that the RD control function be extrapolated within the randomization interval, as it is directly observed.

Multiple-cutoff RD (MCRD) Designs

Another method to increase the generality of RD estimates uses multiple cutoffs along the assignment variable (Angrist & Lavy, 1999; Berk & de Leeuw, 1999; Black et al., 2007; Canton & Blom, 2004; Chay, McEwan, & Urquiola, 2005; der Klaauw, 2002, 2008; Dobkin & Ferreira, 2010; Eggers et al., 2015; Goodman, 2008; Hjalmarsson, 2009; Klasnja & Titiunik, 2017). In such designs, cutoffs generally vary by sites (e.g., school) or times (e.g., year). For instance, a college scholarship can be awarded to students based on their state test scores, but the cutoff score can vary by school district or state (see Figure 5). Multiple-cutoff RD (MCRD) designs support causal generalization by allowing the estimation of a weighted, average RD treatment effect across multiple cutoffs in the assignment variable continuum. In addition, the MCRD approach can also be used to estimate a local RD treatment effect at each cutoff, possibly revealing heterogeneity of treatment effects across the assignment variable continuum.

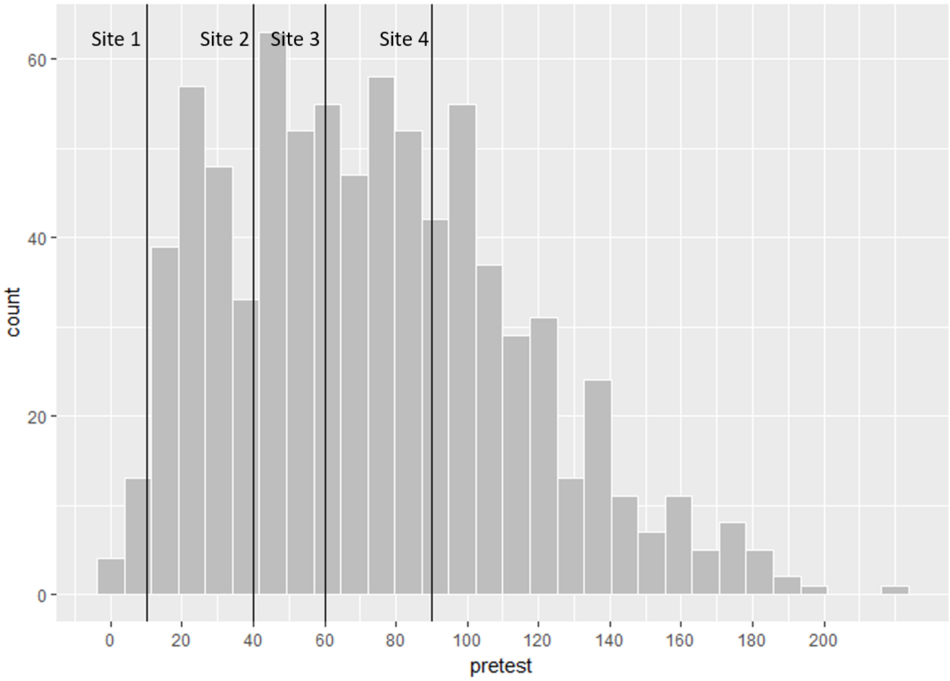


Figure 5. Multiple cutoffs across sites

To date, most studies have used the centering-and-pooling approach to estimate average RD treatment effects across cutoffs in an MCRD design (Cattaneo et al., 2016). Only a few studies have separately estimated a local RD treatment effect at each cutoff in addition to the pooled RD estimate (Canton & Blom, 2004; Cattaneo et al., 2016; der Klaauw, 2002, 2008). The centering-and-pooling approach is implemented by first centering the assignment variable at each cutoff (i.e., normalizing the cutoff) so that all units have a zero cutoff value. Second, observations from all cutoff groups are pooled into a single dataset, and a standard RD is implemented using the pooled dataset consisting of a single cutoff centered at zero (see Figures 5–6). The idea is that the pooled estimates for RD treatment effect can serve as the overall average of the local treatment effects at each cutoff (see Figure 6). Therefore, the pooled RD treatment effect can be considered the average treatment effect across the range of assignment variable continuum where distinct cutpoints are located (Cattaneo et al., 2016).

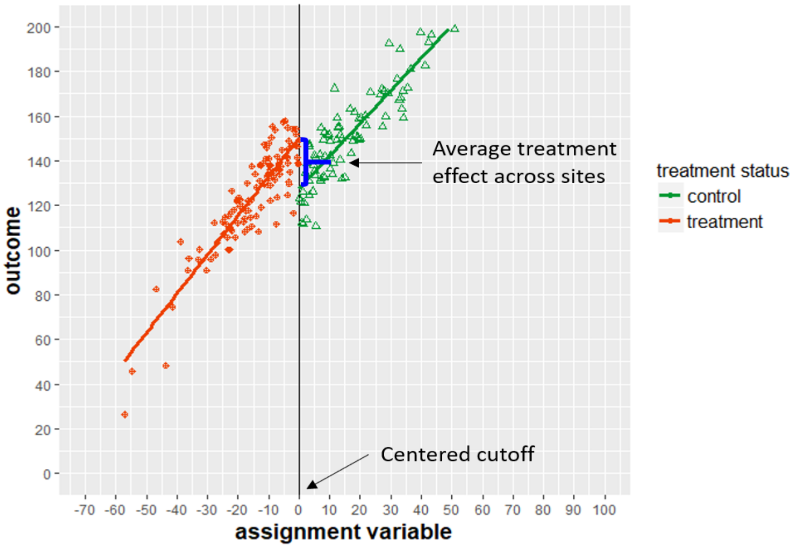


Figure 6. RD estimates in a pooled dataset.

The MCRD approach has been applied in a variety of contexts, including education and public policy, but the most common applications occur in political science (Albouy, 2013; Cattaneo et al., 2016; Eggers et al., 2009; Folke & Snyder, 2012; Hall & Snyder, 2015; Klasnja & Titiunik, 2017; Pettersson-Lidbom, 2008; Uppal, 2009). This is particularly the case in studies where the assignment variable was the impact of winning one election (i.e., becoming an incumbent party) on winning the subsequent election, examined using vote share (i.e., the percentage of total votes a party has secured in an election), and the cutoff was the vote share of the winning party (Butler, 2009; Eggers et al., 2015; Klašnja & Titiunik, 2017; Lee, 2008; Lee, Moretti, & Butler, 2004; Uppal, 2009). The MCRD approach was well-suited for such studies because the vote share of the winning party often varied by electoral constituency (Cattaneo et al., 2016). In light of such variations in cutoffs, almost all political scientists (e.g., Klasnja & Titiunik, 2017) have employed the centering-and-pooling approach by using a margin of victory for the party of interest (i.e., vote share obtained by the party of interest minus the vote share obtained by its strongest opponent) as the assignment variable (e.g., Klasnja & Titiunik, 2017). When using the margin of victory to normalize cutoffs across the electoral constituency, the vote share of the party of interest is converted to relative standing against its strongest opponent. The normalized assignment-variable distribution has a cutoff at zero, at which the party of interest has the same vote share as its strongest opponent. Therefore, by using the margin-of-victory scores (i.e., assignment variable), researchers can then pool all their observations, implement a standard RD analysis with a single cutoff, and estimate a single average treatment effect across cutoffs.

In public-policy research, Black et al. (2007) examined the impact of reemployment services on earnings using cutoffs that varied by time (e.g., weeks) and geography (e.g., local offices). Based on a profiling variable (i.e., expected duration of benefit receipt computed as a function of the individual's characteristics and the state of the local economy), participants were assigned to either the treatment or control condition: Those with higher profiling scores were assigned to treatment, those with lower profiling scores were assigned to the control condition, and those with moderate profiling scores were randomly assigned to either the treatment or control condition. To analyze RD estimates across time and geography (i.e., sites), Black et al. (2007) adopted a two-dimensional RD approach—estimating treatment effects in each dimension separately—as well as the normalizing-and-pooling approach. For example, for the geography dimension, the authors first centered the assignment score (i.e., profiling score) at each site cutoff, then pooled the cases from the cutoff groups into a single dataset and estimated the average treatment effect across the entire assignment variable continuum. Their findings show an unbiased treatment effect estimated in both dimensions, supporting the validity of the MCRD design.

To date, a handful of studies in education research have applied MCRD designs (Angrist & Lavy, 1999; Canton & Blom, 2004; Chay & McEwan, & Urquiola, 2005; Dobkin & Ferreira, 2010; Goodman, 2008; Hoxby, 2000; Kane, 2003; Urquiola, 2006; Urquiola & Verhoogen, 2009; Van der Klaauw, 2002, 2008). For example, Chay, McEwan, and Urquiola (2005) examined the effects of school aid funds (i.e., Chile's 900 School Program) on 4th grade students' gain scores in language and math. The school aid funds were assigned to schools based on their combined mean scores for math and language on a national achievement test. Because the school aid cutoffs varied across Chile's 13 regions, the authors also estimated the pooled RD treatment effect: an average treatment effect across cutoffs. Specifically, they centered the assignment variable—the average mean score for math and language—at each cutoff by creating an average score relative to the cutoff at each region (i.e., subtracting the cutoff score from each school's 1998 average score). Then, they pooled all observations from different regions (i.e., cutoff groups) in a single dataset to run an RD analysis with a single cutoff and estimated an average effect of school aid on student achievement across the 13 regions. Similarly, Dobkin and Ferreira (2010) used an MCRD design to study the effects of school entry law on adult education attainment and job market outcomes (e.g., wages, probability of employment). The treatment in this study was the enforcement of a school entry law, and the assignment variable was age. The cutoff was the date when the school entry law was in effect when the participants were five years old, which varied by research site (i.e., California and Texas) because each state had a different age cutoff for school entry. The authors centered the assignment variable using the number of days from the individual's birthday to the age cutoff date of each state and ran a standard RD analysis with a centered single cutoff to estimate an average single RD treatment effect across the two states.

A small number of studies across disciplines, however, have estimated local RD treatment effects at each cutoff in addition to average RD treatment effects. These studies have used the normalizing and pooling approach to examine heterogeneity in treatment effects depending on cutoff locations on the assignment variable continuum (Canton & Blom, 2004; Cattaneo et al., 2016; Van der Claauw, 2002, 2008). For example, Cattaneo et al. (2016) demonstrated the estimation of both pooled and local RD treatment effects using three empirical examples from political science and education. For the education example, the authors used data from the previously mentioned study evaluating the effects of a Chilean school improvement intervention (P-900) on schools' mean language and mathematics achievement score gains (Chay et al., 2005). In this example, the authors estimated both a single average RD treatment effect and RD effects at six cutoffs determined by region. The results revealed moderate heterogeneity in the effects of P-900 on school mean language gains across cutoffs. Despite the positive average effect of P-900 on language outcomes, the authors also found negative effects at two local cutoffs. These findings suggest that the estimation of the average RD treatment effect using the normalizing and pooling method may "miss the opportunity to uncover key observable heterogeneity in RD design" (Cattaneo et al., 2016, p. 1230). In a study on the effects on college enrollment of merit-based financial aid offered to high school students based on their high school GPAs, Van der Klaauw (2002) also identified differences between local

RD estimates at three different GPA cutoffs and the pooled RD estimate at a centered cutoff.

Taken together, in previous MCRD designs, researchers have mostly estimated a single pooled RD treatment effect. This approach is useful in that it summarizes the weighted average treatment effect across cutoffs, thereby increasing the generality of RD treatment effect estimates beyond a single cutoff point. However, the use of the pooling method alone may ignore potential heterogeneity in treatment effects, given the location of the cutoff on the assignment variable continuum (Cattaneo et al., 2016). The identification of heterogeneous treatment effects could reveal rich information about the treatment effects that may be obscured when averaged across units. For example, information on heterogeneous treatment effects could help identify subgroups of students along the assignment variable continuum for whom an intervention is particularly effective. It should be cautioned, however, that the estimation of local RD treatment effects at each cutoff may not be always feasible or plausible given a particular dataset or research context. A small number of cases around each cutoff, for example, will make it challenging for researchers to detect local RD treatment effects due to low statistical power. In such cases, estimating an average RD treatment effect may be the better option.

In this regard, it is notable that there has been little investigation of estimation approaches for MCRD. Many past studies have applied MCRD designs using the normalizing and pooling approach without explaining the rationale for choosing this analytic method. Given this gap in the current MCRD literature, the *validity* of the causal inferences in past MCRD designs and the *contexts* in which these methods could make credible causal inferences remain unknown. Furthermore, there has been a lack of effort to investigate other analytic approaches for MCRD designs or to explore different strategies in terms of either the internal and external validity of causal estimates or the resulting policy implications.

Kindergarten Math Intervention

Kindergarten mathematics has received increased attention given its critical impact on the development of mathematics understanding in subsequent grades (Clarke et al., 2016; Gersten, Jordan, Flojo, 2005; Jordan & Dyson, 2016; Kohli, Sullivan, Sadeh, Zopluoglu, 2015; Locuniak & Jordan, 2008; Morgan, Farkas, Hillemeier, & Maczuga, 2016; Morgan, Farkas, & Wu, 2009). Data from the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K) revealed that students with mathematics difficulties in kindergarten displayed consistently lower growth rates than their not-at-risk peers in grades 1 through 5 (Morgan, Farkas, & Wu, 2011). The consistently low math gains observed over time for students with math learning disabilities (MLD) in kindergarten widened the fifth grade math achievement gap between these students and students who did not display MLD in kindergarten (Morgan et al., 2011). This study also showed that students who experienced MLD at the end of kindergarten were far more likely than their not-at-risk peers to persistently display MLD throughout elementary and middle school and that kindergarten mathematics achievement is a stronger predictor of MLD than such variables as cognitive delays. In response to the findings of the ECLS-K, there have been focused efforts to screen students with MLD early on and prevent MLD by strengthening the core kindergarten math curriculum and providing additional small-group interventions through multi-tier systems of support (MTSS), such as response to intervention (RTI).

Although there are variations across early math interventions, many focus on the development of number sense (Berch, 2005; Clarke, Baker, & Fien, 2008; Dyson, Jordan, & Glutting, 2013; Gersten & Chard, 1999; National Mathematics Advisory Panel, 2008) using evidence-based instructional practices that have been shown to be effective for at-risk learners (Archer & Hughes, 2011; Baker, Gersten, & Lee, 2002; Coyne, Kame'enui, & Carnine, 2011; Gersten et al., 2009; Kroesbergen & Van Luit, 2003).

Number sense is the ability to connect mathematical concepts to numerical relationships (Gersten & Chad, 1999), which allows students to count, discriminate and coordinate quantities, estimate quantities, discern number patterns, and perform simple number transformations through addition and subtraction (Jordan et al., 2006). Research has showed that most children develop number sense before beginning elementary education (Ginsburg & Golbeck, 2004; Ginsburg & Russell, 1981; Huttenlocher, Jordan, & Levine, 1994; Jordan, Huttenlocher, & Levine, 1994) and that a failure to develop this foundational number concept prior to formal education is associated with difficulties in accessing more advanced mathematics taught in later grades (Jordan et al., 2013; Wu, 1999).

Emerging evidence supports the efficacy of kindergarten math interventions focusing on number sense development. For example, Dyson and Jordan (2011) and Jordan and colleagues (2012) provided 30-minute-long number sense interventions three times a week for eight weeks in a small-group setting that focused on developing the whole number concept in relation to counting, comparing, and manipulating sets to low-income kindergarteners at risk for mathematics difficulties. An evaluation of both interventions revealed that the treatment group made greater gains in both immediate and delayed post-test number sense relative to the business-as-usual (BAU) group (Jordan, Glutting, Dyson, Hassinger-Das, & Irwin, 2012). These results suggest that the continued effects of number sense intervention in kindergarten could help narrow the math achievement gap between students with MLD and students without MLD as they progress in school.

Response to Intervention (RTI)

Many of the research-based kindergarten math interventions are provided through a multi-tier system of support (MTSS) or response to intervention (RTI) (Fuchs & Vaughn, 2012) frameworks. RTI is a multi-tiered approach for the early identification of students at risk for learning difficulties or disabilities, and provides increasingly intensive and focused academic intervention based on student need. In the RTI framework, students' response to intervention and their rate of learning are assessed and monitored through universal screening and progress-monitoring tools several times a year. Although multiple RTI models (Fuchs & Fuchs, 2006) exist, a three-tier model is most commonly used (Fuchs & Vaughn, 2012); in this model, students who do not show an adequate level of learning progress receive a more intensive intervention. Within Tier 1, all students receive scientifically validated instruction in a general classroom setting to ensure that their difficulties are not due to insufficient classroom instruction (Fuchs & Fuchs, 2017). Then, those students who do not adequately respond to the core instruction are provided with supplemental small-group instruction (Tier 2). Within Tier 2, small groups of students (fewer than 6) with similar learning needs receive explicit and systematic instruction three to five times a week for at least 20 minutes per day. If the students in Tier 2 do not meet grade-level expectations, as measured by a progress-monitoring assessment, they are provided individualized instruction that targets their skill deficits (Tier 3). Those students who continue to demonstrate inadequate response to the Tier 3 intervention are then considered for eligibility for special education services (IDEA, 2004). As a prevention system, RTI serves to identify early stage learning difficulties and reduce special education referrals based on the assumption that struggling students are identified early enough so that supplemental intensive instruction can improve student learning outcomes. In what follows, I introduce one of the Tier-2 kindergarten math interventions—ROOTS—to provide the context of this study.

ROOTS

ROOTS is one of the evidence-based Tier-2 kindergarten interventions focused on the development of number sense and whole number concepts, including (a) counting and cardinality, (b) number operations, and (c) base 10/place value (Clarke, Doabler, Fien, Baker & Smolkowski, 2012). ROOTS provides 50 lessons to small groups of 5 students struggling with developing proficiency with whole number concepts and skills for 20 minutes a day, 5 days per week for approximately 10 weeks. In addition to the focused whole number content, ROOTS features the use of the following four validated explicit and systematic mathematics instructional practices: (a) teacher modeling, (b) deliberate practice, (c) visual representations of mathematics, and (d) academic feedback. The ROOTS intervention also facilitates structured opportunities for students to deeply engage in important whole number concepts and skills by having students verbalize their mathematical thinking and discuss their problem-solving methods (Clarke et al., 2016). The efficacy of ROOTS was evaluated using a randomized controlled trial (RCT) design (Clarke et al., 2016; Doabler et al., 2017). Specifically, the 10 lowest children on a screening measure in each of 120 classrooms were randomly assigned to either the treatment or the control condition, meaning that each participating classroom had a different cutoff for screening students into the treatment condition.

The use of ROOTS data in this study is appropriate for the following reasons. First, the multiple cutoffs used to determine ROOTS-eligible students across classrooms enables the application of the MCRD design, in which an average treatment effect across cutoffs is estimated and potential heterogeneity in treatment effect across classrooms can be explored. Second, the use of the RCT design in the original ROOTS study allows the evaluation of the performance of the MCRD design compared to the RCT design. Specifically, the bias and precision of the MCRD design relative to the RCT can be assessed by comparing the treatment effect and power estimates from MCRD with those from RCT.

Predictors of Math Intervention Impact Variability

While research on the development and validation of early mathematics intervention programs is rapidly expanding, there is a growing recognition that not all students respond to these interventions equally (Fuchs et al., 2006; Fuchs & Vaughn, 2012; Starkey & Klein, 2008). The identification of persistently low-achieving students who are not responsive to interventions has motivated research on the factors that predict this non-responsiveness (Miller, Vaughn, & Freund, 2014). Although empirical evidence is emerging and the findings are mixed, these studies have focused on initial math skill (Clarke et al., 2017, in press; Fien et al., 2016; Fuchs, Sterba, Fuchs, & Malone, 2016; Toll & Van Luit, 2013).

Initial math skill. A growing number of studies have investigated initial math skill as a key predictor of the variability of math intervention impact (Clarke et al., 2017, in press; Fien et al., 2016; Fuchs, Sterba, Fuchs, & Malone, 2016; Toll & Van Luit, 2013). Fien et al. (2016) tested the efficacy of an evidence-based math game intervention that focuses on the development of whole number concepts with 250 first-grade students by using an RCT design. The researchers found a statistically significant main effect of treatment on some of their outcome measures, but they did not identify a moderation effect of student initial math skill level on treatment impact. Toll and Van Luit (2013) provided a math intervention focused on basic numerical concepts and math-related language to kindergarteners with low numeracy skills and evaluated if the effect of the intervention was different for students with moderately low skills and those with low initial numeracy skills. The results showed that the intervention was only effective for students with moderately low initial numeracy skills. Fuchs, Sterba, Fuchs, and Malone (2016) found that pre-intervention calculation skill was not predictive of differential responsiveness to a fraction intervention delivered to fourth-grade students displaying low math achievement. However, Fuchs et al. (2016) found that fraction word-problem intervention was more effective for students with greater initial reasoning ability. In addition, Clarke et al. (in press), in their preliminary evaluation study of a small-group kindergarten math intervention program focused on developing whole number concepts and skills, found greater intervention effects for students with lower initial math skills. Taken together, the current literature on the moderating effect of initial math skills on treatment effect has been mixed in terms of the presence of the effect and the direction of impact—whether the intervention is more effective for students with greater or lower initial math skills.

The Current Study

The review of previous research on the MCRD design shows the gaps in the existing MCRD analytic approaches used to estimate RD treatment effects. The current literature also suggests that researchers could effectively investigate the heterogeneity of treatment effect by utilizing the multilevel modeling approach for the MCRD design, not solely relying on a pooling approach. In addition, the emerging literature on heterogeneity in the impact of math interventions warrants additional studies, which could help clarify the source of mixed findings in the literature.

The current study builds on the existing MCRD design in an attempt to increase causal inference beyond the cutoff. In particular, this study is closely aligned with a study by Cattaneo et al. (2016) in that both studies intend to demonstrate and evaluate the optional estimation methods for the MCRD design. This study also extends extant literature on the predictors of heterogeneity in the impact of early math intervention by investigating the moderating role of initial math skill.

The purpose of this study, therefore, is to (a) demonstrate the application of two treatment effect estimation methods (i.e., a conventional pooling method and a multilevel pooling method) for the MCRD designs using Tier 2 kindergarten math intervention data, (b) examine the extent to which the two methods yield unbiased and precise estimates comparable to those from the RCT design, and (c) investigate the moderating role of a classroom characteristic (i.e., classroom cut-point) on the size of the ROOTS intervention effect.

If correctly modeled and properly implemented, the current demonstration will add to recent literature on improving the generality of RD estimates beyond the cutoff by demonstrating and evaluating the applications of different MCRD analytic approaches. Theoretically, both pooled average RD treatment effect and multilevel pooled RD treatment effect estimated at the centered cutoff will help extend the area of causal inference in RD designs beyond a single cutoff. In addition, the findings of this study will add to the growing literature on heterogeneity in math intervention impact, which will help disentangle the mixed findings in the current literature.

Practically, the current study will help program evaluators and policy makers identify a valid and informative program evaluation model when treatment assignment cutoffs vary by sites or times. In particular, the potential identification of heterogeneous treatment effects across cutoffs will provide more specific information about program effectiveness (e.g., how a math intervention may work differently depending on the students' initial math skill). This understanding will further help researchers, program evaluators, and policy makers design, revise, and evaluate interventions that work for students with differential need. Given this backdrop, the research questions are as follows:

1. On average, do the students assigned to the ROOTS intervention outperform those assigned to the control condition across cutoffs?
2. To what extent do treatment effects vary across cutoffs?
3. To what extent does the pooled average RD estimates and the multilevel pooled RD estimates obtained across cutoffs yield unbiased and precise causal estimates relative to the ROOTS RCT benchmark?

4. To what extent is treatment effect variability predicted by classroom characteristics, such as the cut-point used to screen treated students in each classroom?

CHAPTER III

METHODS

In this section, I describe the methodological procedures used in the present study. A variety of graphical, parametric, and nonparametric analyses were implemented for the MCRD using the ROOTS intervention dataset.

Data Source

The ROOTS math 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) kindergarten mathematics intervention. In Years 1 and 2, the participants comprised four school districts in Oregon: one school district located in the Portland metropolitan area and three located in suburban and rural areas of western Oregon. In Years 2 through 3, two school districts from the metropolitan area of Boston, Massachusetts, participated. Table 1 presents the student demographics in percentage by district.

Table 1

Student Enrollment and Demographics in Percentages by Year and School District

	2012-2013				2013-2014	
	Oregon[a]				Massachusetts[b]	
Student characteristics	1	2	3	4	5	6
Student enrollment (*N*)	5,725	2,736	10,808	38,557	6,118	6,843
Race (%)						
White	69	83	69	48	7.8	80.5
Hispanic	18	11	20	26	82.1	6.7
African American	2	1	2	3	7.1	5.4
Asian	2	2	1	14	1.9	4
Native American	<1	<1	2	1	0	0.1
Native Hawaiian/Pacific Islander	1	-	<1	1	0.3	0.3
More than one race	2	1	7	8	0.8	3
Economically disadvantaged (%)	58	64	69	43	83.4	29.8
English language learners (%)	9	5	10	26	18.9	3.1
Students with disabilities (%)	15	14	15	11	12.7	17.2

Note. [a]Statistics are reported for students in grades K-3; [b]Statistics are reported for student K-12.

From 2012 to 2013, across the four districts in Oregon, student enrollment ranged from 2,736 to 38,557 students. The student bodies were approximately 46 to 92% White, 0 to –74% Hispanic, 0–9% African American, 0–16% Asian American, 0–12% American Indian or Native Alaskan, and 0–15% more than one race, respectively. From 2012–2013, approximately 17–86% of district students received a free or reduced-price lunch, 8–23%

of students received special education services, and 5–68% of district students were English language learners.

From 2013–2014, the two districts in Boston had total enrollments of 6,118 and 6,843 students, respectively. In these districts, the student characteristics were 7.8% and 80.5% White, 82.1% and 6.7% Hispanic, 7.1% and 5.4% African American, 1.9% and 4% Asian American, 0.3% and 0.1% American Indian or Native Alaskan, and 0.8 % and 3.0% more than one race, respectively. Approximately 83.4% and 29.8% of each district received a free or reduced-price lunch, 12.7% and 17.2% of students received special education services, and 18.9% and 17.2% of students were identified as English language learners, respectively.

Analytic Sample

Schools. Table 2 presents the count of districts, schools, classrooms, interventionists, and students in the analytic sample. The analytic sample included 14 schools from four school districts in Oregon and 9 schools from two school districts in the metropolitan area of Boston, Massachusetts. In Oregon, six schools were located in one school district in the Portland metropolitan area, and the remaining schools were located in three suburban or rural school districts. In one of the school districts in Boston, all the participants attended the same school. In the other Boston district, the participants attended eight different schools.

Classrooms. The analytic sample consisted of 51 classrooms from Oregon and 55 classrooms from Boston. On average, the classrooms comprised 21.4 students each (*SD* = 5.14) and provided general kindergarten mathematics instruction five days per week. When classrooms had insufficient numbers of students to form intervention and control groups on their own, “intervention classrooms” were formed by combining two or three classrooms. Across four years of project, a total of 11 combined ROOTS classrooms were created using 24 classrooms.

Interventionists. The analytic sample included 71 interventionists. The interventionists included district-employed instructional assistants, interventionists hired for the ROOTS math intervention study, and certified kindergarten teachers. In the original sample, the majority of the interventionists self-identified as female (94%) and White (76%), with 12% identifying as Hispanic and 12% identifying as another race/ethnicity or declining to respond. Almost all the interventionists (92%) had previous experience providing small-group instruction, and 61% had a bachelor’s degree or higher. About half of the interventionists (57%) had taken an algebra course at the college or graduate level. On average, the interventionists had 10.4 years of teaching experience (*SD* = 8.6), and 22% had a current teaching license or certification. The interventionists underwent two additional five-hour professional development workshops. Furthermore, during the implementation period, all interventionists received instructional support from two to five former educators each year through one to four in-class coaching visits varying by the interventionists’ implementation needs (Clarke et al., 2016; Doabler et al., 2016).

Students. The analytic sample used in this study was drawn from the original ROOTS study, which was conducted using a randomized controlled trial design (Clarke et al., 2012). Specifically, in the original ROOTS study, the 10 lowest performing students on a composite score in each classroom were randomized to (a) a ROOTS–small

(2:1) group, (b) a ROOTS–large (5:1) group, or (c) the no-treatment business-as-usual (BAU) control condition.

The initial sample for this study comprised 3,454 kindergarten children, including (a) those who were eligible for random assignment to either the treatment or the control condition (*n* = 1,253) and (b) those who were ineligible for the random assignment but were assessed on the screening and post-test assessments (*n* = 2,201). The following observations were excluded from the analytic sample: (a) those whose screenings, post-test assessments, and assigned conditions were unavailable (*n* = 691) were omitted because an RD analysis could not be implemented without such information; (b) those who were ineligible for the intervention as a result of their performance on the screening measure (above the cut-off rank order), but who received the intervention (*n* = 108), and those who met the eligibility criteria, but were categorized as the non-eligible sample (*n* = 115), were deleted because these non-compliant observations were expected to bias the RD treatment effect estimate; (c) those from the combined classroom for the ROOTS intervention (*n* = 282) were excluded so as to model and estimate the classroom-level variance using multilevel RD design; (d) those whose demographic information (i.e., age and gender) was not available (*n* = 22) were excluded because the covariate balance could not be tested with these observations; and (e) those in the BAU control condition (*n* = 336) were excluded from the RCT sample so as to create an RD treatment sample. As a result, the analytic sample for this study comprised 1,900 kindergarten students for the 2012–2013 (*n* = 486), 2013–2014 (*n* = 976), and 2014–2015 (*n* = 438) academic years (see Figures 7 and 8). The sample included 963 (50.7%) females, 224 (11.8%) ethnic minority students, 252 (13.3%) English language learners, and 85 (4.5%) students receiving special education services. The average age of the sample was 5.36 years.

Table 2

Count of Districts, Schools, Classrooms, Interventionists, and Students in the Analytic Sample

	District	School	Classroom	Interventionists	Student
Oregon (*N*)	4	14	51	29	965
Boston (*N*)	2	9	55	42	935
Total (*N*)	6	23	106	71	1,900

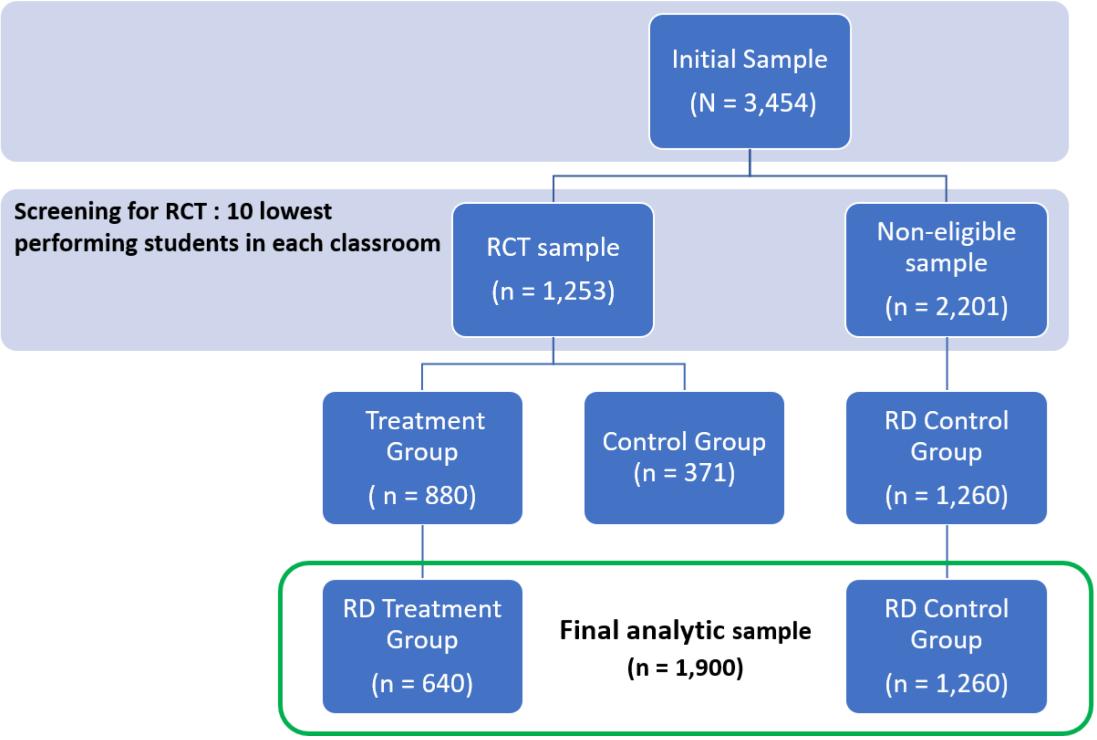


Figure 7. Analytic sample breakdown by condition

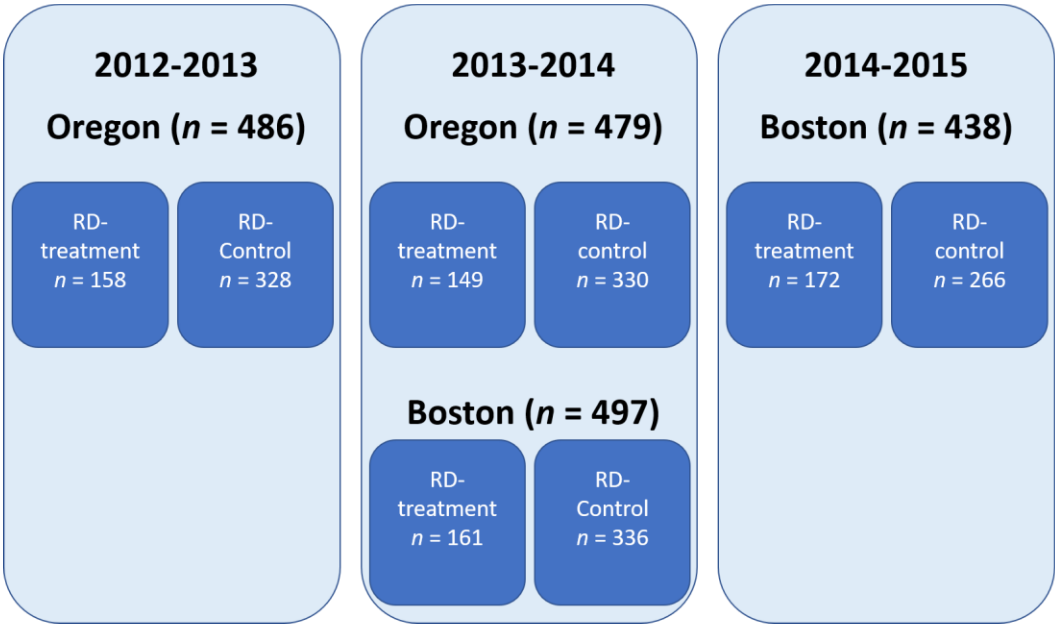


Figure 8. Analytic sample by year and condition.

ROOTS Intervention

ROOTS is a 50-lesson Tier 2 kindergarten intervention program developed to employ evidence-based mathematics instruction to support the development of number sense and whole-number concepts (Clarke et al., 2012). From 2012 to 2015, the intervention was delivered to students randomly selected from the 10 students with the lowest performance on a composite standard score in each participating classroom through a pre-intervention screening.

The composite standard score was formed by combining students' performance on two mathematics proficiency measures: Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke, Rolfhus, Dimino, & Gersten, 2012) and the Number Sense Brief (NSB; Jordan, Glutting, & Ramineni, 2008)[1]. Direct math instruction was delivered to participating students in small groups with either a 2:1 student-

[1] The creation and use of the composite standard score for treatment assignment will be detailed in the next section.

interventionist ratio (i.e., the high-intensity ROOTS group) or a 5:1 student–interventionist ratio (i.e., the low-intensity ROOTS group) in each classroom for 20 minutes five days per week over the 10 weeks of the intervention period. Control students randomly selected from the lowest 10 performing students received no intervention in addition to core instruction. The intervention was delivered to avoid conflicting with students' general kindergarten instruction. The reason for this decision was to prevent disrupting their core instruction. The intervention began in late November to early December and continued through March for each year of the study.

Assignment Criteria

Students were assigned to an intervention condition through the following three steps: First, all students with parental consent were screened on two standardized assessments of early mathematics: ASPENS (Clarke, Rolfhus, Dimino, & Gersten, 2012) and the NSB (Jordan, Glutting, & Ramineni, 2008). Second, students were considered eligible for the 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 (Clarke et al., 2012). The ASPENS and NSB scores of ROOTS-eligible students were separately converted into standard scores, and these two standard scores were then summed to compute an overall composite standard score. Third, the composite standard scores of ROOTS-eligible students were rank-ordered for each classroom, and the lowest 10 students were randomly assigned to one of three conditions: (a) a ROOTS small (2:1) group, (b) a ROOTS large (5:1) group, or (c) a no-treatment BAU control condition (Tier 1 core instruction). After all ROOTS control groups were excluded, the final analytic sample for MCRD design consisted of 106 ROOTS small groups and 106 ROOTS large groups.

Research Design

MCRD. In the present study, I constructed a synthetic RD design from the original RCT dataset. Specifically, I created the RD treatment (RD-T) group by excluding from the RCT data all control group observations below the cutoff in each classroom. Then, I combined the treated observations from two ROOTS treatment groups with different group sizes into one treatment group, given recent findings that there were no significant differences in treatment effects between ROOTS small-group conditions (2:1 and 5:1 student–teacher ratio, respectively) (Clarke et al., 2017). The RD control (RD-C) group comprised the ineligible students who did not meet the screening criteria and did not receive the intervention. Table 3 presents counts and demographic information for the two RD groups identified in this study (*n* = 1,900). 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 treatment assignment and post-test scores were lower in the RD-T group than in the RD-C group. The average ages of the children in the RD-T group and the RD-C group were 5.27 (*SD* = .44) and 5.41 (*SD* = .49), respectively.

Table 3

Student Characteristics and Mean and Standard Deviations of Assignment Scores by Treatment Assignment Condition in the MCRD design

Variables	RD-T (*n* = 640)	RD-C (*n* = 1,260)
Female *n* (%)	334 (52.2)	629 (49.9)
Ethnic Minority *n* (%)	70 (11.0)	154 (12.2)
English Language Learner *n* (%)	124 (19.4)	128 (10.2)
Student with Disabilities *n* (%)	47 (7.3)	38 (3.0)
Age in Years *M* (*SD*)	5.27 (0.44)	5.41 (0.49)
ASPENS[a] Composite Scores *M* (*SD*)	21.87 (17.50)	74.61 (37.46)
NSB[b] Total Raw Scores *M* (*SD*)	12.22 (3.75)	19.86 (4.80)
Composite Standard Scores[c] *M* (*SD*)	-1.52 (0.73)	1.27 (1.40)
Spring SESAT 2[d] Mathematics Scores *M* (*SD*)	460.22 (35.43)	494.29 (37.89)

Note. Total MCRD sample across on conditions, *n* = 1.900. [a]ASPENS =Assessing Student Proficiency in Early Number Sense; [b]NSB = Number Sense Brief; [c]Composite Standard Scores = Composite Standard Scores created from combining standard ASPENSE and NSB scores; [d]SESAT 2 = Stanford Early School Achievement Test 2.

The use of rank order of the composite standard scores in each classroom to determine ROOTS-eligible students (the 10 lowest-performing students) resulted in the current study’s MCRD, such that all ROOTS classrooms used uniquely different cut-off scores. The MCRD design had 106 unique cutoff scores (one for each classroom), ranging from -2.17 to .24 (*M* = -.75, *SD* = .42). As presented in Figure 9, the distribution of the cutoff scores is relatively normal, with many cutoff scores located around the mean cutoff score (between -0.8 and -0.6).

Most of the cutoff values were below 0, indicating that most of the ROOTS-eligible students performed below the mean value of the screening measure. One classroom was found to have fairly lower cutoff scores (-2.17) than others and was considered a mild outlier, given that it was beyond the lower inner fence (1.5 × lower quartile) but within the lower outer fence (3 × lower quartile) (Hoaglin, Iglewicz, & Tukey, 1986). In this study, the outlier was not dropped because it did not significantly alter the results from the main analysis, and the estimation of the treatment effect for the classroom with very low-achieving students was of interest to this study.

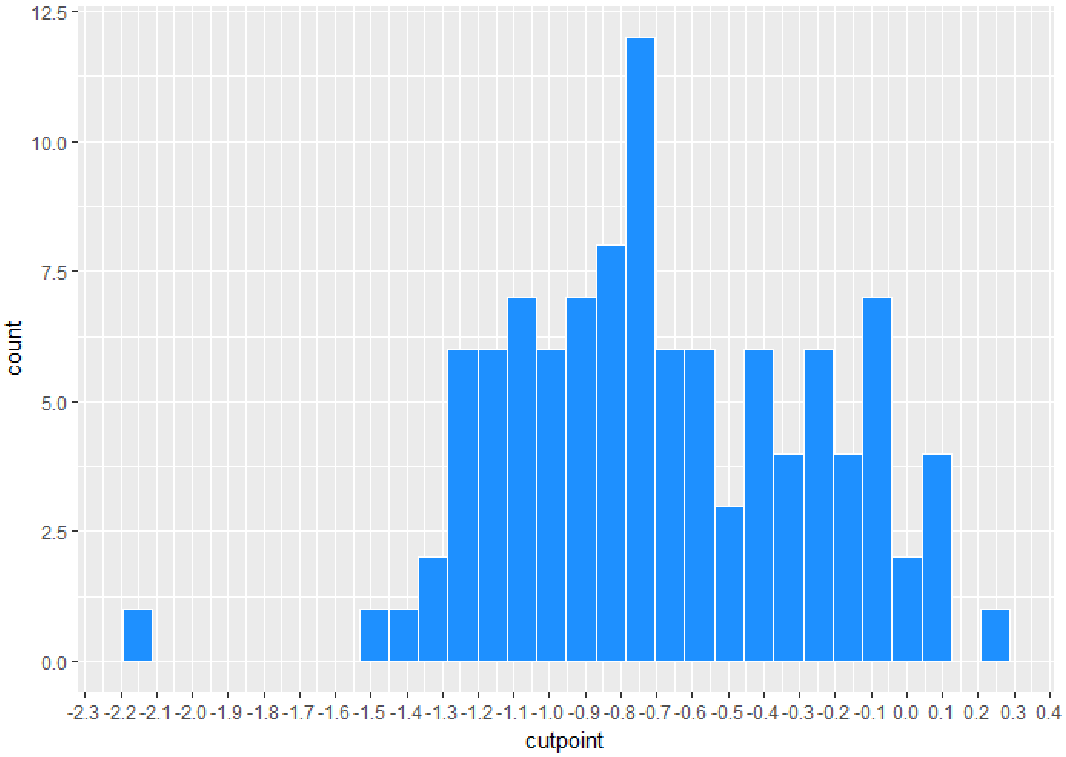


Figure 9. Cut-point distribution in the analytic sample.

Within-study comparison. Within-study comparison (WSC) studies (Cook, Shadish, & Wong, 2008) serve to (a) assess the extent to which quasi-experimental designs are likely to replicate a causal benchmark estimate and (b) identify the contexts and conditions under which quasi-experimental designs produce causal estimates comparable to those from RCTs (Steiner & Wong, 2018). In the WSC design framework, an RCT estimate serves as a causal benchmark, and the treatment effect estimates from non-experimental designs, including RD, difference-in-difference (DID), and interrupted time series (ITS) designs, are compared with those produced by the benchmark design (Cook et al., 2008).

In this study, a WSC study was conducted to evaluate the performance of causal estimates from the MCRD relative to those from RCT design (Cook et al., 2008, LaLonde, 1986). Specifically, I compared the two causal estimates yielded by the MCRD design—(a) a pooled RD estimate and (b) a multilevel pooled RD estimates—and their precision with those from the RCT.

To implement the WSC, both the synthetic RD dataset and the original ROOTS RCT dataset were used. The original RCT dataset provided the benchmark on which the performance of the pooled RD treatment effect estimates yielded from this study's MCRD design were evaluated.

Table 4 presents sample size and descriptive statistics of the full ROOTS RCT. As seen in the table, students assigned to the RCT-T and RCT-C groups were relatively equivalent in terms of age, socioeconomic background, and English language status. There were no statistically significant differences in treatment assignment scores between the RCT treatment and control groups. However, there was a statistical difference in the mean values of one the screening measures (the ASPENS composite) between the RCT treatment and control groups (*F* (1, 885) = 10.86, *p* = .001, Hedges *g* = .23), indicating that the two groups were not entirely comparable on this measure. Therefore, it should be noted that the imbalance in the pretest ASPENS scores can undermine the internal validity of the RCT causal estimate.

The RCT-T group had greater mean value for the post-test measure (Stanford Early School Achievement Test 2 [SESAT 2]) than the RCT-T group, indicating that treated students outperformed control students on an early math achievement measure after the intervention (*F* (1,894) = 25.48, *p* < .001, Hedges *g* = .37).

Table 4

Student Characteristics and Mean and Standard Deviations of Assignment Scores by Treatment Assignment Condition in the RCT Design

Variables	RCT-T (*n* = 639)	RCT-C (*n* =257)
Female *n* (%)	334 (52.3)	138 (53.7)
Ethnic Minority *n* (%)	70 (11)	29 (11.3)
English Language Learner *n* (%)	124 (19.4)	52 (20.2)
Student with Disabilities *n* (%)	47 (7.4)	21 (8.2)
Age in Years *M* (*SD*)	5.27 (.44)	5.26 (.44)
ASPENS[a] Composite Scores *M* (*SD*)	21.86 (17.51)	17.70 (15.75)
NSB[b] Total Raw Scores *M* (*SD*)	12.23 (3.76)	11.45 (3.20)
Composite Standard Scores[c] *M* (*SD*)	-1.52 (.73)	-1.78 (.72)
Spring SESAT 2[d] Mathematics Scores *M* (*SD*)	460.23 (35.46)	447.19 (33.77)

Note. Total RCT sample across on conditions, *n* = 896. [a]ASPENS =Assessing Student Proficiency in Early Number Sense; [b]NSB = Number Sense Brief; [c]Composite Standard Scores = Composite Standard Scores created from combining standard ASPENSE and NSB scores; [d]SESAT 2 = Stanford Early School Achievement Test 2.

Measures

Two standardized assessments of early mathematics were used to screen students into the ROOTS intervention conditions: ASPENS (Clarke, Rolfhus, Dimino, & Gersten, 2012) and the NSB (Jordan, Glutting, & Ramineni, 2008). Students’ ASPENS and NSB scores were separately converted into standard scores, and the two standard scores were then summed to compute an overall composite standard score.

Assessing Student Proficiency in Early Number Sense (ASPENS; Clarke et al., 2012). ASPENS is one of the two screening measures that comprise the assignment variable (i.e., overall composite score) for this study. ASPENS is designed to screen and monitor the progress of students who are at risk for mathematics difficulty in Grades K through 1. ASPENS comprises three curriculum-based measures that assess students' early numeracy proficiency, including number identification, magnitude comparison, and missing numbers. ASPENS is individually administered, taking one minute to administer each subtest. Total raw scores are the number of correct items across the three subtests. ASPENS composite scores are obtained by combining the weighted scores from the three subtests. ASPENS performance categories are also available to show whether students' scores fall below or above a benchmark goal (i.e., "intensive," "strategic," or "benchmark"). On the ASPENS measures, test–retest reliabilities ranged from .74 to .85. The reported predictive validity assessed by the correlation between the fall scores on the kindergarten ASPENS and spring scores on the TerraNova 3 ranged from .45 to .52 (Clarke et al., 2012).

Number Sense Brief (NSB; Jordan et al., 2008). The other screening measure included in the assignment variable is the NSB, which is designed to screen students at risk for later mathematics difficulties. The NSB assesses students' numerical competencies, including counting knowledge and principles, number recognition, number comparisons, non-verbal calculation, story problems, and number combinations. The NSB comprises 33 individually administered items. The total score is the number of correct items. The reported alpha reliability coefficient was .84 for Grade 1 (Doabler et al., 2016). The reported predictive validity assessed by the correlation between the NSB scores in Grades K through 1 and the spring scores on the Woodcock–Johnson Achievement Test in math in Grade 3 was *r* = .62 to .64, and the discriminant validity with spring scores on the Test of Word Reading Efficiency (Torgesen, Wagner, & Rashotte, 1999) in Grade 3 ranged from *r* = .29 to .40 (Jordan et al., 2008).

Stanford Early School Achievement Test 2 Mathematics (SESAT 2; Harcourt Brace Educational Measurement 2003). SESAT 2 Mathematics is the kindergarten version of the Stanford Achievement Test 10th edition series (SAT-10), and it serves as a post-test measure for the current study. SESAT 2 Mathematics is a group-administered, standardized, norm-referenced achievement test with a single subtest. SESAT 2 Mathematics comprises 40 items and takes approximately 40 minutes to administer. The total score is the number of correct items. Student performance level is also made available to identify whether students' scores are below, at, or above average. In one study, the internal consistency of SESAT 2 Mathematics was α = .88 (Doabler et al., 2016); in another, the reported convergent validity assessed by the correlation with Stanford 9 was *r* = .70 to .80 (Carney, 2005). All participating students, including students who did not meet the criteria on the screening measures, were administered the SESAT 2 during the posttest.

Analytic Procedure

Two different versions of MCRD treatment effects were estimated and their performance relative to the RCT estimate evaluated: (a) a pooled RD treatment effect and (b) a multilevel pooled RD treatment effect. Although the analytic procedures implemented for the two estimates present redundancies, I have split the analytic procedure section based on the two MCRD estimates for clarity. For the within-study comparison, the RCT treatment effect was also estimated.

Pooled RD treatment effect. A pooled RD treatment effect can be considered a weighted average treatment effect observed for all treated groups across the assignment variable continuum (Cattaeno et al., 2016). In the following, I demonstrate the analytic procedure used to estimate the pooled RD treatment effect of the ROOTS intervention.

Model assumptions**.** To validate the pooled RD estimate, model assumptions were tested using graphical, parametric, and non-parametric analyses (Bloom, 2012; Lee & Lemieux, 2010). Examination of RD assumptions is important. First, RD designs may involve cases in which the researchers lack control over the treatment assignment, such that “manipulating” scores around the cutoff might occur (Lee & Lemieux, 2010; Skovron & Titiunik, 2015). For example, if college students have knowledge of the treatment assignment process for merit-based scholarships and are able to change or choose their assignment score so that they can be selected into a desired treatment status on the basis of effort, their scores could be different from those immediately below the cutoff (Lee & Lemieux, 2010). When there is sorting around the cutoff, there is a discontinuity in the density at the cutoff. Therefore, to establish the validity of an RD design, it is critical to provide empirical evidence of the continuity in the assignment and outcome variable relationship at the cutoff (Lee & Lemieux, 2010; Skovron & Titiunik, 2015). Additionally, it is important to rule out other plausible alternative explanations for an observed discontinuity in the assignment and outcome variable relationship at the cutoff.

Therefore, I examined whether there was continuity in the covariate distribution (gender, age in years, English-language learner, and special education status) by treatment and control groups at the centered cutoff. This was accomplished by calculating the standardized mean difference in the covariate distribution by treatment and control groups and the variance ratio for covariates at the centered cutoff.

Specifically, before calculating the standardized mean difference in the covariate distribution, equal-width bins on the assignment variable were created. The optimal bin size was calculated using the McCrary density test (McCrary, 2008), as shown below:

$$\hat{b} = 2\hat{s}n^{-\frac{1}{2}} \qquad (3)$$

where $\hat{b}$ is the estimated bin size, $\hat{s}$ is the sample standard deviation of the assignment variable, and *n* is the number of observations. Then, the weighted mean difference in the demographic variable distribution (e.g., English-language learner) was computed within the two equal-width bins around the centered cutoff. The variance ratio was calculated as the mean ratio of the variance of a demographic variable in the treatment group to the variance of the variable in the control group.

Graphical analysis using the non-parametric RD model was also implemented to supplement the covariate balance statistics. The graphical analysis was accomplished by running non-parametric local linear regression (LLR) models where the treatment assignment status predicted demographic covariates (Loader, 1999). In the LLR models, I fit a series of regressions within narrow bandwidths—the width of a window—to allow for non-linearities in the overall function and plotted discontinuities in the regression line representing the relationship between treatment assignment status and demographic covariates.

Model assumptions regarding the manipulation of the assignment variable were not tested because the RD design created from the original ROOTS dataset fits "sharp RD," such that treatment assignment and receipt were completely determined by the value of the assignment variable. Specifically, I built the synthetic RD design from the original RCT dataset by deleting control cases from the treated side of the cutoff (i.e., below the cutoff) and treated cases from the control side of the cutoff (i.e., above the cutoff). The creation of RD data after establishing a cutoff ensured that those who scored above the cutoff had no chance of receiving the treatment; thus, there was no chance to manipulate the assignment variable.

Average treatment effect at the centered cutoff. An overall average of the ROOTS treatment effect (ATE) across cutoffs was estimated. To facilitate this estimation, I first centered the assignment variable at each cutoff so that all units (i.e., classrooms) had a zero cutoff. Then, I pooled observations from all units into a single dataset. Finally, I estimated the ATE at the centered cutoff with the pooled dataset. The pooled RD estimate was computed by weighting number observations in each classroom. All analyses were implemented in R (R Development Core Team, 2012). Following recent recommendations for RD analyses, the effect of assignment at the centered cutoff was examined using graphical analysis and a series of parametric and non-parametric regression analyses (Bloom, 2012; Lee & Lemieux, 2010; Schochet, 2008).

Parametric analysis was implemented using a backward elimination regression method (Cappelleri & Trochim, 1994). Specifically, the outcome (i.e., SESAT 2 Mathematics scores) was regressed on linear, quadratic, and cubic terms of the centered assignment variable and the interaction terms between the dichotomous treatment assignment variable and centered assignment variable. Then, the 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. The general form of the parametric pooled RD model is specified in Equation 1:

$$Y_i = \beta_0 + \beta_1 (x_{\mathrm{i}} - C_j) + \beta_2 (\mathrm{Z_i}) + \beta_3 (x_{\mathrm{i}}\mathrm{Z_i}) + \beta_4 [(x_{\mathrm{i}} - C_j)^2) + \beta_5 ([(x_{\mathrm{i}} - C_j)^2\mathrm{Z_i}) + \beta_6 [(x_{\mathrm{i}} - C_j)^2] + \beta_7 [(x_{\mathrm{i}} - C_j)^3\mathrm{Z_i}) + r_i \quad (1)$$

where Y_i is the SESAT 2 Mathematics score, x_{i} is the composite standard score for intervention assignment, C_j is the cutoff for classroom j, $\mathrm{Z_i}$ is the dichotomous ROOTS intervention indicator, $x_{\mathrm{i}}\mathrm{Z_i}$ is the interaction between composite standard score for intervention assignment and ROOTS intervention indicator, and r_i is the residual term.

A nonparametric analysis was also conducted to complement the parametric estimation (Lee & Lemieux, 2010). Unlike parametric methods, non-parametric methods do not specify a particular functional form in advance; rather, they approximate unknown regression functions from the data. Thus, nonparametric methods allow flexibility in modeling the functional form between assignment and outcome variable along the distribution of the assignment variable. In addition, nonparametric methods are robust to outliers and useful in analyzing categorical data (Bloom, 2012; Whitley & Ball, 2002). In this study, the nonparametric analysis was conducted by implementing local linear regression (LLR) with triangular kernel weights[2] (Loader, 1999). The non-parametric pooled RD model was defined as follows:

$$Y = \alpha_l + \tau Z + \beta_l (x_{\mathrm{i}} - C) + (\beta_r - \beta_l) Z + \varepsilon,$$

$$C - h \leq x_{\mathrm{i}} \leq C + h \quad (2)$$

[2] The triangular kernel assigns the largest weights to the observations in the middle of bin.

where α_l is the intercept from the model on the left-hand side of the cutoff (dichotomous ROOTS intervention indicator), τ is the slope relating the treatment indicator to the outcome, x_i is the composite standard score for intervention assignment, C is the centered cutoff, β_l is the slope relating the assignment variable to the outcome from the regression model on the left-hand side of the cutoff, β_r is the slope relating the assignment variable to the outcome from the regression model on the right-hand side of the cutoff, ε is the residual term, and h is the bandwidth around the cutoff.

Nonparametric analyses were conducted by first identifying an optimal bandwidth, a width of a window in which a series of regressions was fitted using the Imbens and Kalyanaraman (IK) algorithm (Imbens & Kalyanaraman, 2011). In identifying the optimal bandwidth, it is important to balance the bias and precision associated with wider and smaller bandwidths. Wide bandwidths may produce biased impact estimates of conditional predicted means, but have greater precision because they use a greater number of observations for estimation. Narrow bandwidths may produce less biased impact estimates, but have less precision because they use fewer observations. After the optimal bandwidth was identified, the assignment variable continuum was segmented into smaller equal-width bins. The optimal bin size was calculated using the McCrary density test (McCrary, 2008), as shown below:

$$\hat{b} = 2\hat{s}n^{-\frac{1}{2}} \tag{3}$$

where $\hat{b}$ is the estimated bin size, $\hat{s}$ is the sample standard deviation of the assignment variable, and n is the number of observations. Then, a series of regressions across bins within the optimal bandwidth was fit around the cutoff. To assess the sensitivity of nonparametric estimates to the bandwidth choice, the nonparametric analysis was repeated using bandwidths half and twice the optimal size as comparators. Standard errors for the nonparametric estimates were obtained using bootstrapping techniques (with 1,000 repetitions).

Multilevel RD treatment effect. A two-level hierarchical linear model (HLM; Raudenbush & Bryk, 2002) was then used to estimate the weighted multilevel treatment effect and examine potential heterogeneity in the effects of the ROOTS intervention across classrooms. The multilevel model partitioned variance and covariance into discrete levels of data structure (i.e., the student and classroom levels), which allowed a test of whether student-level treatment effects varied by classroom. Full-information maximum likelihood estimation was used to estimate the models. All analyses in this portion of study were implemented in R (R Development Core Team, 2012).

As with the traditional pooled ATE at the centered cutoff, the assignment variable (i.e., composite standard scores) was centered at each cutoff so that all units had a zero cutoff and the RD treatment effect at the centered cutoff could be estimated. Then, an unconditional two-level model was specified to estimate the mean post-test SESAT mathematics score and the amount of student and classroom variation in students' post-test SESAT mathematics score components (see Equations 4 and 5).

Level 1 (students): $Y_{ij} = \beta_{0j} + r_{ij}$ (4)

Level 2 (classrooms): $\beta_{0j} = \gamma_{00} + u_{0j}$ (5)

In Equations 4 and 5, Y_{ij} is the posttest SESAT mathematics scores in classroom j for student i, β_{0j} is the mean posttest SESAT Mathematics score for classroom j, γ_{00} is the mean posttest SESAT mathematics score across classrooms, r_{ij} is the student level residual term, and u_{0j} is the classroom level residual term relating to mean posttest SESAT mathematics score.

Then, a conditional RD model was specified by adding level-1 predictors. Specifically, the conditional RD model was defined as follows:

Level 1 (students): $$Y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - C_j) + \beta_{2j}(Z_{ij}) + \beta_3(x_{ij} - C_j)^2 + r_{ij} \tag{6}$$

Level 2 (classrooms):

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30} \tag{7}$$

where Y_{ij} is the post-test SESAT Mathematics score for student i in classroom j, x_{ij} is the composite standard score (assignment variable) 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, β_{0j} is the mean SESAT mathematics score for classroom j, β_{1j} is the slope relating the assignment score to the post-test SESAT mathematics score in classroom j, β_{2j} is the slope relating treatment receipt to the post-test SESAT mathematics score in classroom j, β_{3j} is the slope relating the assignment score squared term to the post-test SESAT mathematics score in classroom j, r_{ij} is the student-level residual term, and the u's are the classroom-level residual terms. The assignment score variable was allowed to vary at level 2 as the random effect term for this variable significantly improved the model fit. The slope relating treatment receipt to post-test SESAT mathematics scores in classroom j was not allowed to vary at level 2, as the random effect term for this variable did not significantly improve the model fit.

RCT treatment effect. A two-level HLM was built to estimate the average treatment effect with the RCT dataset (Raudenbush & Bryk, 2002). As with the RD analysis, the two-level HLM was used to account for common variance among students (level 1) nested within the same classrooms (level 2). Full-information maximum likelihood estimation was used to estimate the model. An unconditional model was first specified to estimate the mean post-test SESAT mathematics score across classrooms and the amount of student and classroom variation in students' SESAT mathematics score components. The unconditional model was defined as follows:

Level 1 (students): $Y_{ij} = \beta_{0j} + r_{ij}$ (8)

Level 2 (classrooms): $\beta_{0j} = \gamma_{00} + u_{0j}$ (9)

where Y_{ij} is the post-test SESAT mathematics score for student *i* in classroom *j, β_{0j}* is the mean post-test SESAT mathematics score for classroom *j*, γ_{00} is the mean post-test SESAT mathematics scores across classrooms, r_{ij} is the student-level residual term, and u_{0j} is the classroom-level residual term relating to the mean post-test SESAT mathematics score.

Then, a conditional model was built to estimate the RCT treatment effect, in which students' post-test SESAT mathematics scores were regressed on a dummy-coded treatment indicator (i.e., assigned to the treatment condition or not). To improve the precision of the RCT estimate, the assignment variable (i.e., composite standard score) was included in the model. The final conditional model of SESAT outcomes was specified as follows:

Level 1 (students): $Y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij}) + \beta_{2j}(\mathrm{Z}_{ij}) + r_{ij}$ (10)

Level 2 (classrooms):

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

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

$$\beta_{2j} = \gamma_{20} \tag{11}$$

where Y_{ij} is the post-test SESAT mathematics score for student i in classroom j, x_{ij} is the composite standard score (assignment variable) for student i in classroom j, Z_{ij} is the dichotomous ROOTS intervention indicator for student i in classroom j, β_{0j} is the mean post-test SESAT mathematics score for classroom j, β_{1j} is the slope relating the assignment score to the post-test SESAT mathematics score in classroom j, β_{2j} is the slope relating treatment assignment to the post-test SESAT mathematics score in classroom j, r_{ij} is the student-level residual term, and u_{0j} is the classroom-level residual term relating to the mean post-test SESAT mathematics score. The assignment score variable, the slope relating treatment receipt to the post-test SESAT mathematics score in classroom j were not allowed to vary at level 2, as the random effect terms for these variables did not significantly improve the model fit.

Within-study comparison. The MCRD design yielded two causal estimates: (a) a pooled RD estimate and (b) a multilevel RD estimate. These estimates were then compared with those from the RCT in terms of bias and precision.

Bias. Bias in the pooled RD estimate and the multilevel pooled RD estimate drawn from the MCRD was evaluated by computing the difference between the RCT causal estimates of the treatment effect and the MCRD estimates at the centered cutoff. Specifically, the standardized bias in each MCRD estimate compared to the RCT benchmark was calculated as the difference between the MCRD estimate and the treatment effect estimate from the RCT design. The standardized bias of the pooled RD estimate was estimated as below:

$$SB_{Pooled\,RD} = (\pi_{Pooled\,RD} - \pi_{RCT}) \times \frac{1}{\sigma_{RCT}} \tag{12}$$

where $SB_{Pooled\,RD}$ is the standardized bias of the pooled RD estimate, $\pi_{Pooled\,RD}$ is the pooled RD estimate of the treatment effect, π_{RCT} is the treatment effect estimate produced by the RCT, and σ_{RCT} is the standard deviation of outcome (i.e., SESAT mathematics score) observed in the RCT. Standardized bias assessed using this equation indicates the magnitude of the bias of the MCRD estimate in standard deviation units. Following the criterion used by Tang et al. (2017), a RCT/MCRD difference less than .10 standard deviation units as used to determine the robustness of the estimates of the two MCRD estimates (pooled RD and multilevel RD estimates) compared to the RCT causal estimate.

Power. The precision of the two MCRD treatment effect estimates (i.e., pooled RD and multilevel RD estimates) was examined by comparing the standard errors (*SE*) of each MCRD estimate at the centered cutoff with those of the RCT causal estimate. The *SE*s were calculated taking into account the different sample sizes across the two designs (i.e., MCRD and RCT). Following the method used in previous studies, the SEs were equated by the difference in the sample size between the RCT and MCRD designs (Schochet, 2009; Tang et al., 2017). More specifically, if the sample size for the MCRD design is n_{MCRD} and that for the RCT design is n_{RCT}, then the SE for the MCRD estimate is proportional to $\sqrt{1/n_{MCRD}}$ and that for the RCT is proportional to $\sqrt{1/n_{RCT}}$ (Schochet, 2009). Then, a fair comparison of SEs in the RCT and MCRD can be made by comparing the SE of the MCRD design to $\sqrt{n_{RCT}/n_{MCRD}}$ multiplied by the SE of the RCT (Tang et al., 2017). Therefore, the adjusted *SE* for RCT was estimated as follows:

$$SE_{\text{adjusted\_RCT}} = (\sqrt{n_{RCT}/n_{MCRD}}) * SE_{\text{RCT}} \qquad (13)$$

Exploring the Predictors of Treatment Impact Heterogeneity

Following the estimation of the multilevel RD model, I explored a predictor of treatment impact heterogeneity across classrooms using classroom characteristics: the cut-point used to screen treated students. In addition to partitioning variance and covariance into discrete levels of data structure, multilevel modeling allows the modeling of cross-level interaction terms, thereby enabling an examination of how varying cutoff values chosen for each classroom might be associated with unit treatment effects—the degree to which the treatment assignment cutoff amplifies or attenuates post-test outcomes. Specifically, a two-level RD model was specified by adding level 1 and level 2 predictors. The conditional RD model was defined as follows:

Level 1 (students): $Y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij\,-}\,C_j) + \beta_{2j}(\mathrm{Z}_{ij}) + \beta_3\{(x_{ij\,-}\,C_j)^2 + r_{ij}$ (14)

Level 2 (classrooms):

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21*}C_j$$

$$\beta_{3j} = \gamma_{30} \qquad (15)$$

where Y_{ij} is the post-test SESAT Mathematics score for student *i* in classroom *j*, x_{ij} is the composite standard score (assignment variable) 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*, $cutoff_j$ is the cutoff value for classroom *j*, β_{0j} is the mean SESAT mathematics score for classroom j, β_{1j} is the slope relating the assignment score to the post-test SESAT mathematics score in classroom j, β_{2j} is the slope relating treatment receipt to the post-test SESAT mathematics scores in classroom j, β_{3j} is the slope relating the assignment score squared term to the post-test SESAT mathematics score in classroom j, r_{ij} is the student-level residual term, and the u's are the classroom-level residual terms.

Note that in Equation 15, the cutoff value for classroom j, C_j was added as a classroom-level predictor. Specifically, the treatment effect estimate in classroom j was modeled as a function of variation in the cutoff value chosen for each classroom. Therefore, at level 2, the term γ_{21} represents the relationship between the location of the cutoff value and the treatment–outcome relationship in classroom j. The assignment score variable was freed to randomly vary across classrooms. The slope relating treatment receipt to the post-test SESAT mathematics score in classroom j, and the slope relating the assignment score squared to the post-test SESAT mathematics score in classroom j were not allowed to vary at level 2, as the random effect terms for these variables did not significantly improve the model fit.

CHAPTER IV

RESULTS

In this chapter, I describe the results of the graphical, parametric, and nonparametric RD analyses implemented using the ROOTS intervention dataset. The results of the RD model assumption tests are presented first. Then, the results of two versions of RD analyses (i.e., pooling RD analysis and multilevel pooling RD analysis), RCT estimation, and within-study comparison follow.

Model Assumptions

Discontinuity at the cutoff of assignment variable continuum. Figure 10 demonstrates the relationship between the probability of being assigned to the treatment condition and the composite standard score. As expected for a synthetic RD, the probability of being assigned to the treatment condition is 1.0 below the zero cutoff on the composite standard assignment variable continuum, while the probability of being assigned to the treatment condition is 0 above the cutoff.

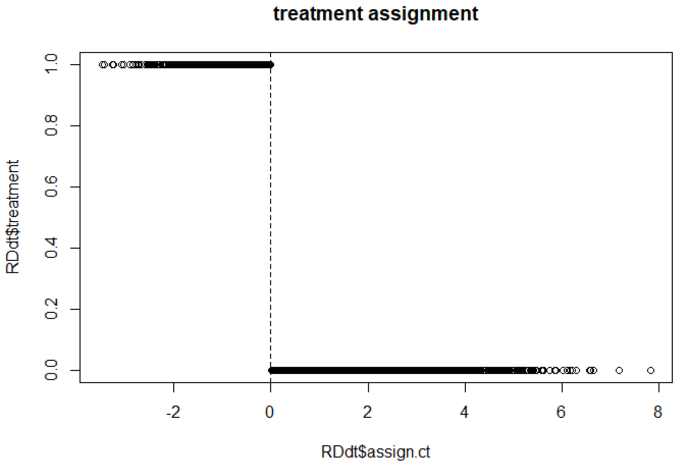


Figure 10. Treatment assignment as a function of the assignment variable score.

Covariate balance at the centered cutoff. Table 5 presents the weighted mean difference in a demographic variable distribution along with the mean ratio of the variance of a demographic variable in the treatment group to the variance of that variable in the control group. Results showed that there was no statistically significant discontinuity in covariates, including gender, ethnicity, limited English proficiency, and special education, at the cutoff. These results, along with those from graphical analyses (see Appendix), indicate that there was a balanced distribution of covariates across RD treatment and RD control group members at the centered cutoff.

Table 5

Covariate Balance Statistics

Covariates	∆ Difference	*SE*	*p*	Variance ratio
Female	-0.04	0.04	0.31	1.01
Student age	-0.05	0.03	0.14	0.92
Limited English proficiency	0.03	0.03	0.39	1.10
Special education	0.01	0.02	0.67	1.14

Note. ∆ 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 control group.

Pooled MCRD Models

Parametric model. Table 6 provides the pooled RD estimates of the ROOTS intervention from both the parametric and nonparametric MCRD models. The implementation of the backward elimination regression method revealed that a model with a linear assignment variable, a quadratic assignment variable, and a treatment indicator served as the best fitting model.

Results showed that the assignment variable (i.e., composite standard scores) was associated with student posttest SESAT mathematics scores ($\beta_{1\,=}$ 14.35, p < 0.001), indicating that students with higher assignment scores had higher posttest scores than those with initially lower assignment scores (see Figure 11). The estimate of the pooled RD treatment effect was statistically significant ($\beta_{2,}$ = 5.19, , p < 0.001). The quadratic assignment variable was not statistically related to the posttest SESAT outcome (β_3 = -0.21, p > 0.05).

Table 6

Parametric Pooled RD Estimates of SESAT Mathematics Outcome

	Estimate	*SE*	*t*
Intercept	467.00	0.83	564.04^{***}
Composite standard scores	14.35	0.69	20.93^{***}
Composite standard scores squared	-0.21	0.23	-0.95
Treatment	5.19	1.32	3.94^{***}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

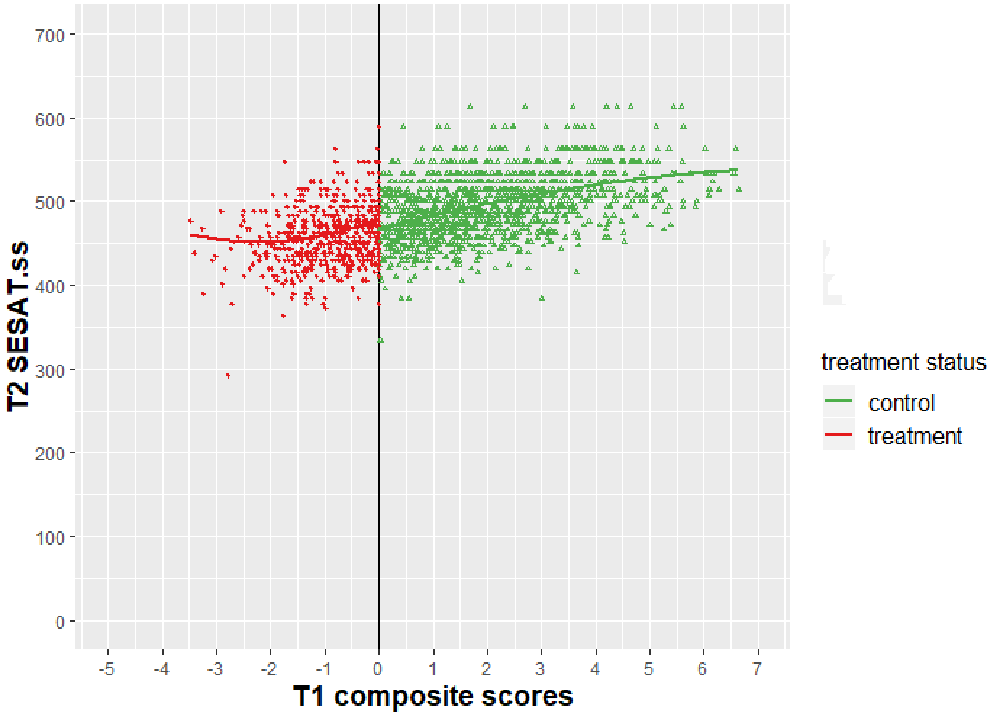


Figure 11. Parametric plot of the assignment–outcome variable relationship by treatment assignment status.

Nonparametric model. Figure 12 depicts the nonparametric local linear regression plot for the assignment—outcome variable relationship using the traditional pooling MCRD method. As a supplementary specification for the parametric model, a nonparametric method using a local linear regression was implemented within the optimal bandwidth (2 points) around the cutoff chosen from the IK procedure, meaning a series of regressions were fit 2 points below and above the cutoff. The bins were equally sized at 0.08 points. By restricting the analysis to observations within this bandwidth, the local linear analysis used 39.4% of the student observations. As seen in Table 7**,** the nonparametric pooled RD estimate was 6.63, which was statistically significant ($p < 0.05$). This result indicates that, on average, treated students scored 6.63 points higher in posttest SESAT mathematics than control students at the centered cutoff. The results of the sensitivity analysis showed that the pooled RD estimate with half-sized and double-sized bandwidths was 8.81 ($p < 0.05$) and 4.17 points ($p > 0.05$), respectively, indicating that the nonparametric impact estimate was sensitive to the size of the bandwidth. That is, this result suggests that the narrow bandwidth tends to produce an estimate that is closer to (i.e., less biased than) the one from the optimal bandwidth (6.63) because observations far away are excluded from the estimation. However, the resulting estimate is less precise due to the smaller number of observations contained in the narrower bandwidth (*SE* = 5.03). This result also suggests that the wide bandwidth, in contrast, yielded a more biased estimate (4.17) because the observations far away were included in the estimation. However, the resulting estimate tends to having more precision (*SE* = 2.96) because more observations within the wider bandwidth were used for the estimation. Taken together, the magnitude of the pooled parametric RD estimate was smaller than that of the pooled nonparametric RD estimate with the optimal bandwidth (6.04 point). The nonparametric impact estimate with optimal bandwidth resulted in higher standard errors than the parametric impact estimate.

Table 7

Nonparametric Pooled RD Estimates of SESAT Mathematics Outcome

	Half bandwidth		Optimal bandwidth		Double bandwidth	
	Estimate	*t*	Estimate	*t*	Estimate	*t*
Intercept	465.21 (2.75)	169.08***	466.22 (1.96)	237.67***	467.55 (1.67)	280.40***
Composite standard scores	20.43 (5.26)	3.89***	15.72 (2.03)	7.74***	13.63 (1.05)	13.01***
Treatment	8.81 (3.83)	2.30*	6.63 (2.96)	2.24*	4.17 (2.55)	1.63

Note. Nonparametric pooled RD estimates are shown with standard errors in parenthesis. Nonparametric estimates have bootstrapped standard errors (repetition = 1,000).
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

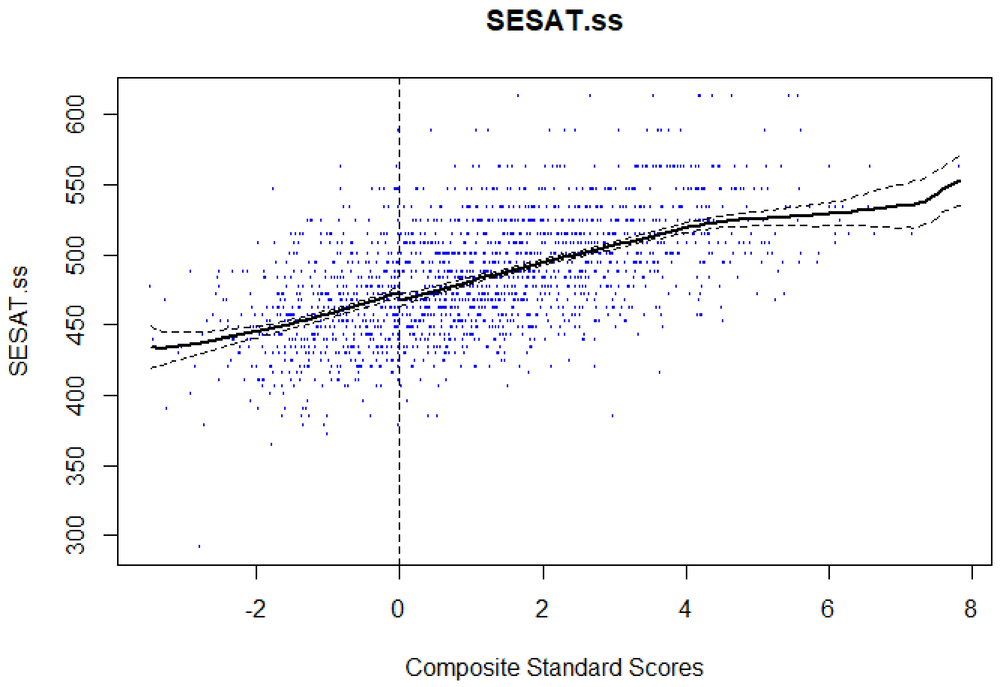


Figure 12. Nonparametric plot of the assignment–outcome variable relationship by treatment assignment status.

Multilevel Pooling MCRD Model

Table 8 presents the estimates for the unconditional SESAT outcome model. The mean posttest SESAT mathematics score across classrooms was 483.50 (*SD* = 40.35). The intraclass correlation coefficient (ICC) showed that a total of 23.29% of the variation in student posttest SESAT mathematics scores was explained by classroom-to-classroom differences. Variance component estimates indicated that mean student posttest SESAT mathematics scores varied significantly between classrooms (r_{0j} = 381.00, p < 0.001).

Table 8

Two-Level Unconditional SESAT Mathematics Outcome Model

Fixed effect	Estimate	*SE*	*t*
Average SESAT mean, γ_{00}	480.35	2.07	233.4***
Random effect	Variance component	SD	
Student (level-1), r_{ij}	1,255.00***	35.43	
Mean SESAT score (level-2),u_{0j}	381.00***	19.52	
ICC ROOTS classroom	.23		

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 9 presents the estimates for the final two-level conditional model. Results of the final model revealed that, at the student-level (level 1), the group-centered assignment variable (i.e., composite standard scores) was associated with student posttest SESAT mathematics scores ($\gamma_{10} = 15.47$, $p < 0.001$). Students with higher assignment scores had higher posttest outcomes than those with initially lower assignment scores. The quadratic assignment variable was associated with student posttest SESAT mathematics score ($\gamma_{30} = -0.57$, $p < 0.01$), indicating that the slope relating assignment variable and student posttest SESAT mathematics score significantly decelerated as the assignment variable value increased. The estimate of the multilevel pooled RD treatment effect was statistically significant ($p < 0.05$), indicating that, on average, treated students scored 5.62 points higher on the SESAT mathematics posttest than control students at the centered cutoff.

.

Table 9

Two-Level Conditional SESAT Mathematics Outcome Model

	Estimate	*SE*	*t*
Fixed effect			
Intercept, γ_{00}	488.09	2.49	187.93***
Inter-student predictors			
Composite standard scores, γ_{10}	15.47	0.95	16.30***
Treatment, γ_{20}	5.62	2.31	2.43*
Composite standard scores squared, γ_{30}	-0.57	0.18	-3.18**
Random effect	Variance component	*SD*	
Student (level 1), r_{ij}	671.78***	25.92	
Mean SESAT score, u_{0j}	465.20***	21.57	
Composite standard scores, u_{1j}	7.75***	2.78	
ICC ROOTS classroom	.41		

Note. *SE* = Standard error.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

RCT Estimation

Table 10 reports the RCT treatment effect estimates from the final two-level conditional model of the SESAT outcome. The mean posttest SESAT mathematics score across classrooms was 450.82 (*SD* = 35.28). The ICC for the unconditional model showed that a total of 39.97% of the variation in student posttest SESAT mathematics scores was explained by classroom differences. Variance component estimates indicated that mean student posttest SESAT scores varied significantly between classrooms (r_{0j} = 423.90, $p < 0.001$).

Results of the final conditional model revealed that at the student level (level 1), the group-mean centered assignment variable was associated with the student posttest SESAT mathematics scores (19.18, $p < 0.001$). This result indicates that students with higher assignment scores had stronger posttest SESAT mathematics performance than those with initially lower assignment scores. The RCT treatment effect estimate was 8.11 points and was statistically significant ($p < 0.05$), indicating that treated students, on average, scored 8.11 points higher than control students on the posttest.

Table 10

RCT Treatment Effect Estimates

	Estimate	*SE*	*t*
Fixed effects			
Intercept, γ_{00}	450.82	2.56	173.83***
Composite standard scores, γ_{10}	19.18	1.25	15.29***
Treatment, γ_{20}	8.11	1.93	4.25**
Random effect	Variance component	*SD*	
Student (level 1), r_{ij}	636.70***	25.23	
Mean SESAT score (level 2), u_{0j}	423.91***	20.59	
ICC ROOTS classroom	0.40		

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Within-study Comparison

Table 11 compares the performance of the MCRD design with the RCT with respect to bias and precision. Based on the criteria for evaluating the performance of causal estimates from the MCRD relative to those from the RCT, both MCRD estimation methods—the pooling approach and the multilevel approach—produced a standardized bias of less than 0.10 SD. This result means the differences between the two MCRD estimates and the RCT benchmark were within the guidelines for interpreting the magnitude of the standardized mean difference as small (Tang et al., 2017). Therefore, these results indicate that the MCRD design produced unbiased treatment effect estimates compared to the RCT causal estimates, regardless of the estimation method. As for precision, the MCRD design, regardless of the estimation method, produced larger standard errors than those from the RCT design, confirming that the RCT design provides more precise estimates.

Table 11

Results of Within-Study Comparison

	Estimate	*SE*	*t*	Bias
RCT benchmark	8.11	1.31[a]	4.25^{***}	
Pooled MCRD				
Quadratic regression	5.19	1.32	3.94^{***}	-0.08
Local linear regression with half BW	8.81	5.03	2.30^{*}	-0.02
Local linear regression with optimal BW	6.63	2.96	2.24^{*}	-0.04
Local linear regression with double BW	4.17	2.55	1.63	-0.11
Multilevel pooled MCRD	5.62	2.31	2.43^{*}	-0.07

Note. a = Standard errors of the RCT estimate were adjusted given the different sample sizes across the two designs (i.e., $N_{\mathrm{RD}} = 1200$, $N_{\mathrm{RCT}} = 896$).
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Predictors of Treatment Impact Heterogeneity

Lastly, I added a classroom-level predictor to model the heterogeneity in treatment effects across classroom units. Results of the final conditional model revealed that there was a statistically significant average treatment effect at the student level (γ_{01} = 11.63, $p < 0.01$). The cut-point used to screen students into treatment conditions across classrooms statistically interacted with the treatment effect parameter (γ_{21} = 7.43, $p < 0.05$) (see Table 12). Figure 13 provides a graphical presentation of the treatment effect variability between the classrooms for the post-test SESAT mathematics score. The treatment effect estimates varied widely across classrooms, depending on the cutoff value used to screen students into the treatment conditions in each classroom, with an approximate range between -5 and +12 points. Specifically, a higher cutoff value chosen to screen students into treatment conditions was associated with greater treatment effect. It was particularly notable that the classrooms with the lowest cutoff values showed negative treatment effect, indicating that in these classrooms (lower end of cut-point distribution), the students assigned to the ROOTS intervention largely underperformed compared to those assigned to the control condition on the post-test SESAT mathematics.

Table 12

Multilevel Model Results Predicting SESAT Mathematics Outcome from Classroom Cut-point

	Estimate	*SE*	*t*
Fixed effect			
Intercept, γ_{00}	478.84	4.19	114.30^{***}
Inter-student predictors			
Composite standard scores, γ_{10}	15.90	0.96	16.65^{***}
Treatment, γ_{20}	11.63	3.65	3.19^{**}
Composite standard scores squared, γ_{30}	-0.62	0.18	-3.47^{***}
Inter-classroom predictors			
Cut-point, γ_{01}	16.36	5.01	3.27^{**}
Cut-point, γ_{21}	7.43	3.61	2.06^{*}
Random effect	Variance component	*SD*	
Student (level 1), r_{ij}	672.23^{***}	25.93	
Mean SESAT score, u_{0j}	388.03^{***}	19.70	
Composite standard scores, u_{1j}	7.08^{***}	2.66	
ICC ROOTS classroom	0.37		

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

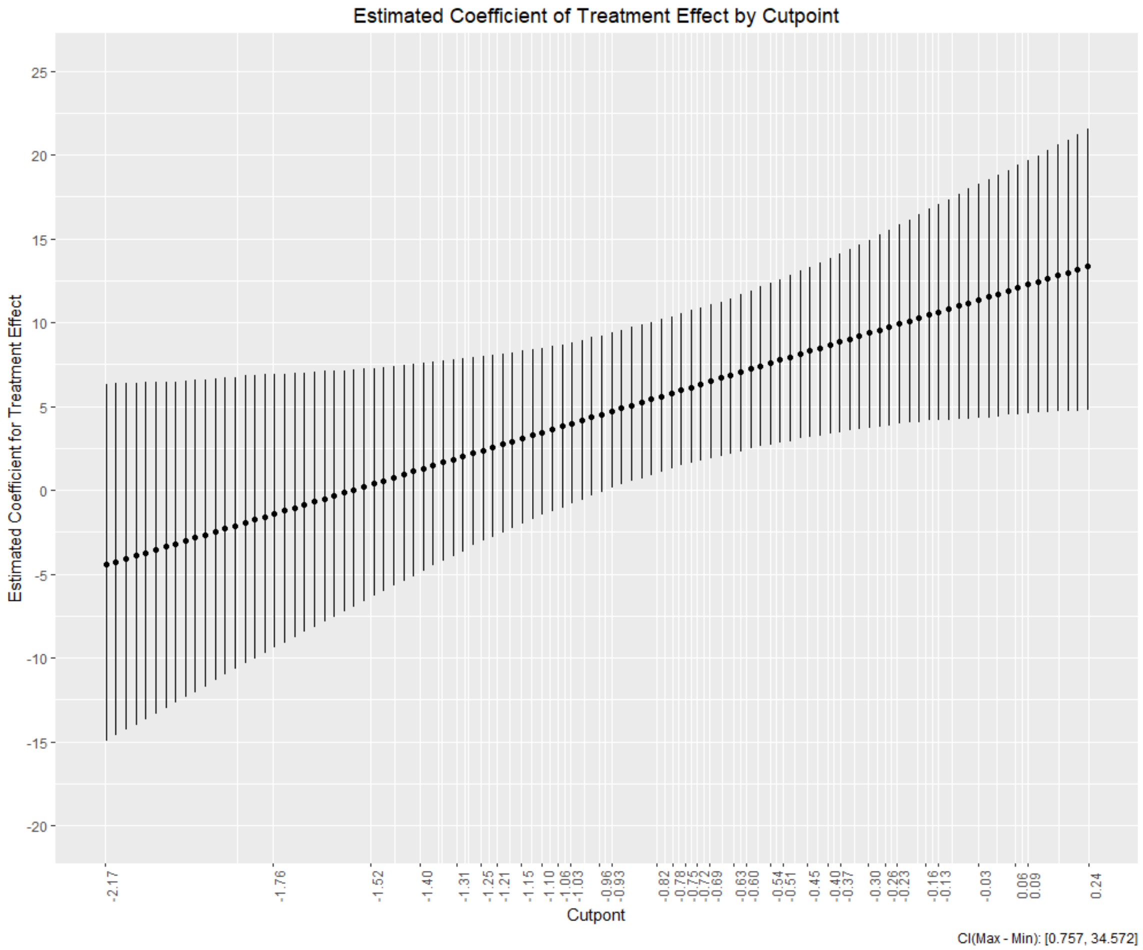


Figure 13. Treatment effect estimates as a function of classroom cut-point. Estimates for each cutoff are displayed with a 95% confidence interval.

CHAPTER V

DISCUSSION

The primary objective of this study was to demonstrate two estimation methods for multiple-cutoff RD (MCRD) designs and to evaluate the two methods with respect to the validity and efficiency of the causal inference using kindergarten math intervention data. Specifically, the intervention effect was estimated using a conventional pooling method and a multilevel pooling method. The bias and power of the resulting MCRD estimates were compared with an RCT benchmark. The secondary objective of this study was to examine if treatment effect heterogeneity was associated with classroom-level characteristics—cut-scores used to screen students into the treatment condition in each classroom.

The findings of this study are as follows. First, at the centered cutoff, treatment students scored higher on the SESAT posttest outcome than control students. Second, all of the MCRD methods produced unbiased treatment effect estimates comparable to a benchmark RCT estimate; however, the power in the MCRD design was lower than in the RCT, regardless of the estimation method. Third, the cut-point used to screen students into the treatment condition moderated the treatment effect, with a greater treatment effect observed in the classrooms with a larger cutoff value. In the following, I discuss the findings as well as implications for program evaluation design theory and practice.

Estimation Methods

Results from the two estimation methods revealed a positive impact of the ROOTS intervention on student math achievement across the treatment cutoffs. Across cutoffs, treatment students had higher scores on the SESAT posttest outcome. The current result is not consistent with the preliminary findings regarding the impact of the ROOTS intervention (Clarke et al., 2016), which showed no treatment effect on gains from fall to spring for the SESAT outcome. The difference in the findings might be due to the difference in the sample. The ROOTS impact studies only used the sample from the first cohort of students of a four-year project, whereas this study used the full sample from all four cohorts. The difference in the finding also might have resulted from the different outcomes used in each study. The prior ROOTS impact studies used the math gains from fall to spring, but this study only used the posttest outcome.

The current results also revealed variations in the treatment effect estimate and statistical power depending on the estimation method. The magnitude of the RD estimates was largest when using the nonparametric pooling method (6.63), followed by the multilevel model (5.62) and the parametric pooled model (5.19). Not surprisingly, the nonparametric pooling method produced the highest standard errors and the parametric pooling method produced the lowest standard errors. More discussion about the robustness and precision of these impact estimates will follow in the next section of this chapter.

The pooling method demonstrated in this study has been widely used in prior MCRD studies to estimate an average treatment effect across cutoffs (Cattaneo et al., 2016; Dobkin & Ferreira, 2010; Gooman, 2008). A relatively small number of studies also have estimated the local treatment effect at each cutoff separately for the MCRD design. The current study is closely in line with these prior studies utilizing the pooling method given that it used the same methods for the MCRD analysis; however, it was also distinguished from these studies in that it employed a multilevel model to account for and model group-level variance and explored if different estimation methods provide similar or different results in addition to the pooling method. As noted earlier, estimation of the local RD treatment effects at each cutoff may not always be feasible and/or plausible. For example, a small number of cases around each cutoff could make it impossible to estimate the local treatment effect at each cutoff. However, when using multilevel modeling, researchers could investigate whether the treatment effect at the individual level varies across treatment groups by testing the random effect term for the treatment effect parameter. Multilevel modeling also enables the exploration of source of the treatment effect variability at both the individual and group levels (e.g., student and classroom characteristics).

Within-study Comparison

The results from the WSC revealed that both MCRD estimation methods produced unbiased treatment effect estimates relative to the RCT benchmark. The RCT benchmark–MCRD estimate difference was less than 0.10 SD units. However, this study found that MCRD estimates were still underpowered compared to those of an RCT design.

Specifically, the nonparametric pooled RD estimate with optimal bandwidth was found to have the smallest standardized bias (0.02 SD units) but was the most imprecise. This finding suggests that the nonparametric pooling method correctly modeled the functional form of the data but lacked statistical power. The multilevel RD estimate was also found to have very small bias (0.07 SD units) and have standard error that is larger than that from the parametric pooled RD model and smaller than that from the multilevel model. The small bias and relatively moderate statistical power of the multilevel RD estimate may be due to the fact that the associations among observations within same groups were accounted for in the multilevel design. The parametric pooled RD estimate was found to have the largest bias (.08 SD); however, it was statistically the most precise among all MCRD estimates, and even as precise as the RCT benchmark. However, interpretation of the high statistical power that was obtained from the parametric pooled model requires caution; the standard error of the regression coefficient could have been underestimated because the clustered structure of the data (i.e., correlations among observations within clusters) was not taken into account in the single-level linear regression model.

Taken together, the nonparametric pooling method appears to provide the most unbiased RD estimate, but given its low statistical power, it is best utilized with a large sample. If a large sample is not available, then the multilevel design could be the next best option for the MCRD analysis, given that it has a very small bias and a relatively high statistical power in comparison to nonparametric methods. If a particular design results in a small sample combined with a limited number of intervention units/sites, the parametric pooled RD method could be the best option. As noted earlier, however, caution should be taken to interpret the results from the parametric pooled RD model due to potential limitations regarding ignoring the clustered structure of the data. Finally, I recommend that analysts and program evaluators employ all three MCRD estimations along with graphical analysis to examine the sensitivity of treatment impact estimates and resulting power estimates, depending on the estimation method, to further probe the source of the different results. Given that each MCRD estimation method demonstrated in this study has its own strength and weakness (i.e., bias-statistical power tradeoff), the optimal estimation method should be determined by considering the research/evaluation context.

A few past studies have also used multiple estimation methods for MCRD designs (Cattaneo et al., 2016; Eggers et al. 2015; Kane, 2003; Van der Klaauw, 2002). These studies estimated the weighted average treatment effect at the centered cutoff and the local treatment effect at different cutoffs separately. For example, Cattaneo et al. (2016) demonstrated the estimation of the average treatment effect at the centered cutoff using the normalizing-and-pooling approach and the estimation of the local treatment effects at each cutoff to the differences in the resulting pooled average RD estimate and the local RD estimates. The current study is closely related to that research in that it also demonstrates the different estimation methods—the pooling method and multilevel modeling—and examined the differences in the resulting estimates. However, the current study is also different from previous studies in that it evaluated the performance of the different MCRD estimates using WSC. In other words, this study tested the internal validity and statistical power of the MCRD estimates compared to those from the RCT in addition to estimating treatment effect.

The findings regarding the performance of MCRD estimates compared to those of the RCT add to the literature on alternative RD designs that were proposed to improve the generality of RD estimates (Tang et al., 2017; Wing & Cook, 2013). Most of the prior studies on alternative RD designs also conducted WSC to evaluate the performance of the RD alternative, allowing evaluation of the internal validity of the design. To date, most of these WSC studies within alternative RD frameworks were conducted with comparative RD designs (Wing & Cook, 2013; Tang et al., 2017). For example, Wing and Cook (2013) demonstrated that, when the comparative regression function (pretest scores) is parallel with the RD control function in the untreated part of the assignment variable continuum, it is possible to extrapolate the RD treatment effect estimate at the cutoff beyond the cutoff (in the treated part of the assignment variable continuum). In addition, the authors also found that CRD estimates were unbiased and precise, and comparable to those from the benchmark RCT used in the WSC.

When findings about performance of CRD and MCRD estimates evaluated by WSC are taken together, it appears that both CRD and MCRD estimates could be used to improve the generality of the treatment effect or extend the area of causal inference beyond the small area around a single cutoff. In addition, both designs seem to produce very small bias, which supports the internal validity of the resulting RD estimates. However, as noted earlier, the MCRD estimates seem to have a lower statistical power than CRD estimates, given that the CRD estimates were found to have a statistical power close to that of an RCT (Tang et al., 2017; Wing & Cook, 2013). In addition, the implementation of MCRD designs that use multilevel modeling typically requires a multitude of research sites/units, which might be burdensome or even implausible for the evaluation of intervention programs that are implemented across a small number of sites or intervention units. It should be noted, however, that the MCRD does not require the onerous parallel regression function assumption and the additional data (pretest scores) that the CRD requires, potentially giving the MCRD more practical utility.

Predictors of Treatment Impact Heterogeneity

The location of the cutoff value chosen for screening the students in each classroom strongly moderated the treatment effect. A greater treatment effect was observed for students from classrooms with higher cutoff values than those from classrooms with lower cutoff values. Remarkably, the treated students from the classrooms with the lowest cutoff values among ROOTS classrooms had *lower* post-test score than those in the control condition by approximately 5 points.

To further probe what the cutoff value could possibly denote, I examined the correlation between the cutoff values and the intervention group's mean pretest scores, which produced a moderate correlation (r = .57, p < .01). This result confirmed that the cutoff value could be interpreted as a proxy of the intervention group's mean initial math skill. Therefore, the cutoff value moderation effect suggests that intervention groups with better initial math skill benefitted more from the ROOTS intervention than those with lower initial math skill. Furthermore, the negative treatment effects for the lowest performing students among all ROOTS participants also suggest that the students with the lowest initial math skill were not responsive to the ROOTS intervention; they continued to struggle with math at the end of the intervention.

The current finding regarding the moderation effect of initial math skill on the size of the treatment effect is inconsistent with a recent study that demonstrated that a game-based math intervention designed to develop students learn whole number concepts was equally effective across the pretest distribution (Fien et al., 2016). One plausible explanation for the inconsistency between these studies might be the difference in the mode of intervention (i.e., game-based vs. group-based intervention). In the game-based intervention, the individualized game play might work equally well for students with different initial math skill by allowing them to master the concepts at their own pace (Nelson, Fien, Doabler, & Clarke, 2016). In the school-based intervention, where multiple students with different levels of need are grouped to receive intervention, however, the extra needs of at-risk students might not be optimally met when they receive the intervention with moderately at-risk peers in the same group.

Notably, the current result opposes the findings of Clarke and colleagues (in press), who reported a greater positive impact of the ROOTS intervention for students with lower initial skill. The discrepancy between these findings is more striking when comparing the direction of the treatment effect that was observed at the lowest end of the distribution of the initial math skill in the two studies. In contrast to Clarke and colleagues' (in press) findings, this study found a negative treatment effect that favored those in the control condition, meaning that the treated students with the lowest initial math skill underperformed their peers in the control condition at the end of intervention.

This contradictory result may be due to differences in the measures of "initial math skill" used in each study. In the present study, initial math skill was measured using a composite standard score formed by combining students' performance on two mathematics proficiency measures (i.e., NSB and APSENS). The NSB and ASPENS are proximal to the intervention, while the Test of Early Mathematics Ability-3 (TEMA-3), used to define initial math skill in Clarke et al. (in press), is distal to the intervention. Therefore, these findings suggest that the nature of a measure used to assess *initial math skill*, for example, how it is operationalized, and whether the measure is proximal or distal to the intervention, may play a critical role in evaluating the differential impacts of math interventions based on students' initial math skill.

Advantages of MCRD Design

Collectively, findings from this study suggest the following advantages of the MCRD design. First, MCRD designs could be used to overcome one of the key limitations of the basic RD design—the limited causal inference at a single cutoff on the assignment variable continuum. In MCRD designs, the use of multiple cutoffs extends the area of causal inference beyond the vicinity of the single cutoff, thereby enabling the estimation of a pooled "average" RD treatment effect across cutoffs. The pooled RD treatment effect estimate then serves to provide a summary of the weighted average treatment effect across all cutoffs when there is no interaction between the assignment variable and the outcome or heterogeneity in the treatment effect. In other words, the estimation and interpretation of an average treatment effect by pooling observations might be meaningful when a constant treatment effect is identified.

Second, the use of the multilevel pooling MCRD method allows the investigation of heterogeneity in treatment effects by intervention sites (e.g., districts, schools). The application of multilevel modeling to MCRD design enables researchers to test whether the student-level treatment effect varies by intervention group by partitioning variance and covariance into discrete levels of data structure (i.e., student- and group-level) and modeling treatment effect variance. Similarly, the ability to model cross-level interactions allows the examination of how site characteristics (e.g., cutoff value, mean SES) moderate treatment effects. Taken together, the use of the multilevel pooling MCRD method appears to offer a great advantage for program evaluation, in that it can identify specific information about treatment effectiveness—how academic interventions work differently for students with different initial skill levels or for students from different contexts. Lastly, MCRD designs maintain ethical research and administrative practices, with respect to providing treatment to those most in need.

Limitations and Future Directions

The current study has several limitations to note in regard to the findings. First, the use of synthetic RD limits the external validity of the study findings. That is, it remains unknown whether the MCRD estimation examined in this study would produce equally unbiased causal estimates when noncompliance exists in practice. Therefore, future studies are needed to evaluate the validity of MCRD methods with fuzzy RD designs.

Second, this study focused on parametric analysis and only uses nonparametric analysis for the traditional pooling MCRD method. Given the advantages of nonparametric analysis (i.e., correct functional form specification), future research should apply nonparametric analysis across all MCRD methods, especially for multilevel MCRD. In this study, nonparametric methods were not applied for the multilevel MCRD because of the small within unit sample size. The use of nonparametric methods for multilevel MCRD method would aid in preventing the identification of false discontinuities, which could result from a mis-specified parametric model. Further, the identification of consistent findings across parametric and nonparametric methods would provide additional support for the validity of multilevel RD estimates.

Third, this study examined a limited number of predictors of treatment effect heterogeneity. Given recent research findings concerning the relationships between individual-, group- and classroom-level factors and math development (e.g., Cragg & Gilmore, 2014; Crosnoe et al., 2010; Hill; Hindman, Skibbe, Miller, & Zimmerman, 2010), future study is warranted to investigate how various classroom- and teacher-factors account for unit-to-unit variability in treatment effects. Findings from such studies could reveal whether classrooms with higher instructional quality or highly qualified teachers are associated with larger treatment effects.

Fourth, the study findings revealed that the MCRD methods had lower statistical power than RCT design regardless of estimation method. Therefore, future research on ways to increase the statistical power of causal estimates in MCRD designs would be useful for improving MCRD designs and increasing their utility in practice.

Finally, the findings of this study—specifically, the performance of the MCRD estimate—lacks external validity. That is, it is unknown whether same results would be found with different populations and in different settings. To establish the external validity of performance of the MCRD design, it is essential to replicate this study with different samples and variables of interest at different time points.

Implications and Conclusions

The findings of this study have several potential theoretical and practical implications. First, the findings of this study fill an important gap in the current literature on alternative RD designs to improve the generality of RD estimate or extend the area of causal inference beyond the small area near a single cutoff. Specifically, the ability to estimate the average treatment effect across multiple cutoffs suggests that MCRD design could improve the causal generalization of RD designs beyond a single cutoff. In addition, this study showed that the use of multilevel modeling for MCRD designs could estimate potential heterogenous treatment effects.. In either scenario, the MCRD design allows for the extension of the area of causal inference along the entire assignment variable continuum.

Second, the identification of the comparability of the causal estimates of MCRD and RCT provides empirical evidence of the internal validity of different MCRD estimation methods, which, in turn, supports the MCRD as a strong alternative to the RCT when individuals are assigned to either to the treatment or control condition using cutoffs that vary by site or time. However, considering the relatively lower statistical power of the MCRD estimates compared to the RCT estimates, MCRD might be most useful for multi-site program evaluations with relatively large within site samples.

Third, the findings of this study concerning the impact of the ROOTS intervention offer a valuable contribution to the growing literature on early mathematics interventions by identifying heterogeneity in the impact of the ROOTS intervention by initial math skill (Clarke et al., in press; Fuchs et al., 2016; Toll & Van Luit, 2013). The current literature on the relationship between initial math skill and intervention response is mixed. The results of this study, which contradict prior findings in the literature, suggest that careful examination of the measure of initial math skills and their relation to the intervention could untangle the intricacy of this body of literature. That is, when different measures were used to measure initial math skills, they could have defined different facets of the "mathematics skill" construct. These measures also differed in their relation to the intervention; some were proximal to the intervention, while others were more distal.

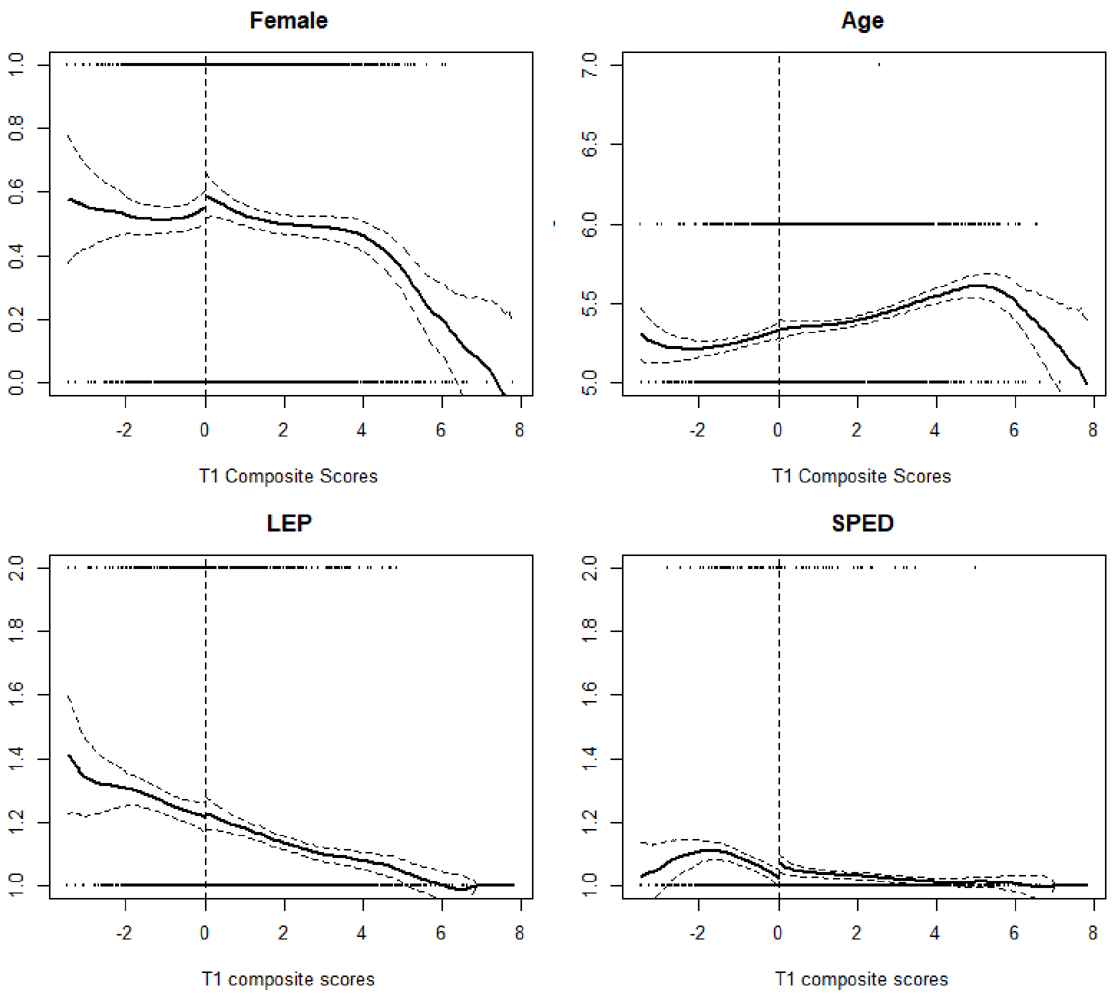
Fourth, the findings suggest that the MCRD could serve as a useful, valid, and informative program evaluation design when varying treatment assignment cutoffs are used by intervention site or group. The ability to identify specific information concerning the RD treatment effect, including possible heterogeneity, may further help program evaluators and policymakers evaluate, modify, and develop interventions that could work for all students with different levels of needs. Importantly, the identification of differential treatment effects will provide intervention developers with valuable information that they can use to enhance their treatment programs. For example, if a math intervention program is seen to have a significantly positive impact for students at moderate risk, but not for those most at-risk, educators and policymakers could revisit their goals and reconsider the intervention's target population or modify the program accordingly. If the aim in providing a math intervention is to help the most struggling students, the program providers can adjust the curricula and instruction or provide more intensive math instruction to the most at-risk students separately.

RD designs have gained popularity for evaluating cutoff-based interventions in education science, given their ability to make robust causal inferences and facilitate ethical delivery of interventions (Jacob et al., 2012). However, one critical limitation of RD designs is that the causal inference is restricted to the vicinity around the cutoff, and the treatment effect beyond the cutoff remains unknown. To address this issue, many efforts have been made to extend the area of causal inference beyond the cutoff in RD designs. Furthermore, under the Individuals with Disabilities Education Act (IDEA, 2004), many research-based interventions have been developed to support students who may be or are at risk of learning disabilities. Evaluations of these academic interventions have revealed that some groups of students persistently do not respond to intervention, which has further motivated researchers to seek a more nuanced picture of when, for whom, and under what circumstances interventions work (Reardon & Stuart, 2017).

The motivation of this study was to examine a method for generalizing the RD estimate and to model treatment effect variability within the RD framework. Although more investigations are needed, the findings of this study suggest that MCRD designs have potential to extend the area of causal inference beyond the vicinity of the single cutoff and identify the processes and mechanisms through which academic interventions are differentially effective across individuals and contexts. The findings of this study may help program evaluators and educators plan rigorous research designs and better target future interventions.

Appendix

Covariate balance at the centered cutoff



REFERENCES CITED

Albouy, D. (2013). Partisan representation in Congress and the geographic distribution of Federal funds. *Review of Economics and Statistics*, *95*(1), 127-141.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects using Instrumental Variables. *Journal of the American Statistical Association*, *91*, 444-472.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton university press.

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. New York, NY: Guilford Press.

Baker, S., Gersten, R., & Lee, D. S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *The Elementary School Journal*, *103*(1), 51-73.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame‘enui, E. J., & Thomas Beck, C. (2008). Reading fluency as a predictor of reading proficiency in low performing high poverty schools. *School Psychology Review*, *37*(1), 18−37.

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38,* 333–339. doi: 10.1177/00222194050380040901

Berk, R. A., & De Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association*, 94(448), 1045-1052.

Black, D., Galdo, J., & Smith, J. A. (2007). *Evaluating the regression discontinuity design using experimental data*. Unpublished paper.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43-82.

Borman, G. D., & Dowling, N. M. (2006). Longitudinal achievement effects of multiyear summer school: Evidence from the Teach Baltimore randomized field trial. *Educational Evaluation and Policy Analysis*, *28*(1), 25-48.

Butler, D. M. (2009). A regression discontinuity design analysis of the incumbency advantage and tenure in the US House. *Electoral Studies*, *28*(1), 123-128.

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance* (No. 14194). Cambridge: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14194.pdf.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2016). *Regression Discontinuity Designs Using Covariates*. Working Paper, University of Michigan.

Campbell, D. T. (1969). Reforms as Experiments. *American Psychologist*, *24*, 409-429.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research. Handbook of research on teaching*. Chicago, IL: Rand McNally.

Canton, E., & Blom, A. (2004). Can student loans improve accessibility to higher education and student performance. *World Bank Policy Research Working Paper,* 3425.

Cappelleri, J. C., & Trochim, W. M. K. (1994). An illustrative statistical analysis of cutoff-based randomized clinical trials. *Journal of clinical epidemiology*, *47*(3), 261-270.

Cappelleri, J. C., & Trochim, W. M. K. (1995). Ethical and scientific features of cutoff-based designs of clinical trials: A simulation study. *Medical Decision Making*, *15*(4), 387-394.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., & Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4), 1229-1248.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, *37*(2), 403-429. https://doi.org/10.1002/pam.22051.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review, 95*(4), 1237-1258.

Chen, M. K., & Shapiro, J. M. (2007). Do harsher prison conditions reduce recidivism? A discontinuity-based approach. *American Law and Economics Review*, *9*(1), 1-29.

Clarke, B., Baker, S. K., Smolkowski, K., Doabler, C. T., Strand Cary, M., & Fien, H. (2015). Investigating the efficacy of a core kindergarten mathematics curriculum to improve student mathematics learning outcomes. *Journal of Research on Educational Effectiveness, 8,* 303–324. doi: 10.1080/19345747.2014.980021

Clarke, B., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012). *A randomized control trial of a Tier 2 kindergarten mathematics intervention*. Eugene: University of Oregon.

Clarke, B., Doabler, C. T., Smolkowski, K., Baker, S. K., Fien, H., & Strand Cary, M. (2016). Examining the efficacy of a Tier 2 kindergarten mathematics intervention. *Journal of learning disabilities*, *49*(2), 152-165.

Clarke, B., Doabler, C. T., Kosty, D., Kurtz Nelson, E., Smolkowski, K., Fien, H., & Turtura, J. (2017). Testing the efficacy of a kindergarten mathematics intervention by small group size. *AERA open, 3*(2)*,* 1–16. doi: 10.1177/2332858417706899

Clarke, B., Doabler, C. T., Smolkowski, K., Turtura, J. Kosty, D., Kurtz Nelson, E., Fien, H., & Baker, S. K. (in press). Exploring the relationship between initial mathematics skill and the impact of a kindergarten mathematics intervention on Student Mathematics Outcomes. *Exceptional Children*.

Clarke, B., Rolfhus, E., Dimino, J., & Gersten, R. M. (2012). *Assessing Student Proficiency of Number Sense (ASPENS)*. Longmont, CO: Cambium Learning Group, Sopris Learning.

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.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within‐study comparisons. *Journal of policy analysis and management*, *27*(4), 724-750.

Coyne, M. D., Simmons, D. C., Hagan-Burke, S., Simmons, L. E., Kwok, O.-M., Kim, M., . . . Rawlinson, D. A. M. (2013). Adjusting beginning reading intervention based on student performance: An experimental evaluation. *Exceptional Children, 80,* 25–44. Retrieved from http://cec.metapress.com/content/F4321275232V11WX

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63-68.

Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., & Clarke-Stewart, K. A. (2010). Instruction, teacher–student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology*, *102*(2), 407-417.

Dobkin, C., & Ferreira, F. (2010). Do school entry laws affect educational attainment and labor market outcomes?. *Economics of education review*, *29*(1), 40-54.

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of learning disabilities*, *46*(2), 166-181.

Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., & Snyder, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, *59*(1), 259-274.

Ferreira, F., & Gyourko, J. (2009). Do political parties matter? Evidence from US cities. *The Quarterly journal of economics*, *124*(1), 399-422.

Fien, H., Doabler, C. T., Nelson, N. J., Kosty, D. B., Clarke, B., & Baker, S. K. (2016). An examination of the promise of the NumberShire level 1 gaming intervention for improving student mathematics outcomes. *Journal of Research on Educational Effectiveness*, *9*(4), 635-661.

Folke, O., & Snyder, J. M. (2012). Gubernatorial midterm slumps. *American Journal of Political Science*, *56*(4), 931-948.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it?. *Reading research quarterly*, *41*(1), 93-99.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Hope, S. K., Hollenbeck, K. N., Capizzi, A. M., ... & Brothers, R. L. (2006). Extending responsiveness-to-intervention to math problem-solving at third grade. *Teaching Exceptional Children*, *38*(4), 59-63.

Fuchs, L. S., Malone, A. S., Schumacher, R. F., Namkung, J., Hamlett, C. L., Jordan, N. C., ... & Changas, P. (2016). Supported self-explaining during fraction intervention. *Journal of Educational Psychology*, *108*(4), 493-508.

Fuchs, L. S., Sterba, S. K., Fuchs, D., & Malone, A. S. (2016). Does evidence-based fractions intervention address the needs of very low-performing students? *Journal of Research on Educational Effectiveness, 9*, 662–677. doi: 10.1080/19345747.2015.1123336

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of learning disabilities*, *45*(3), 195-203.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First Impact Study. Final Report.* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gersten, R. M., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33,* 18–28. doi: 10.1177/002246699903300102

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of learning disabilities*, *38*(4), 293-304.

Ginsburg, H. P., & Baroody, A. J. (2003). Test of Early Mathematics Ability, Third Edition (TEMA-3). Austin, TX: Pro-Ed.

Goodman, J. (2008). Who merits financial aid?: Massachusetts' Adams scholarship. *Journal of public Economics*, *92*(10-11), 2121-2131.

Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations*. Unpublished manuscript.

Hahn, J., Todd, P., & van de Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201-209.

Harcourt Educational Measurement. (2003). *Stanford achievement test [SAT-10]*. San Antonio, TX: Author.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American educational research journal*, *42*(2), 371-406.

Hindman, A. H., Skibbe, L. E., Miller, A., & Zimmerman, M. (2010). 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*, *25*(2), 235-250.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*(396), 991-999.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*(4), 1239-1285.

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *61*(2), 467-476.

Imbens, G. W., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933-959.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615-635.

Individuals with Disabilities Education Act of 2004. 20 U. S. C.§1400 et seq. (2004).

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1), 226-244.

Jacob, R., Zhu, P., Somers, M. A., & Bloom, H. (2012). A Practical Guide to Regression Discontinuity. MDRC.

Jordan, N. C., & Dyson, N. (2016). Catching math problems early: Findings from the number sense intervention project. In *Continuous Issues in Numerical Cognition* (pp. 59-79).

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647.

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowker (Eds.), *Mathematical difficulties: Psychology and intervention* (pp. 45-58). Oxford: Academic Press.

Kane, T. J. (2003). *A quasi-experimental estimate of the impact of financial aid on college-going* (No. w9703). National Bureau of Economic Research.

Klašnja, M., & Titiunik, R. (2017). The incumbency curse: weak parties, term limits, and unfulfilled accountability. *American Political Science Review, 111*(1), 129-148.

Kohli, N., Sullivan, A. L., Sadeh, S., & Zopluoglu, C. (2015). Longitudinal mathematics development of students with learning disabilities and students without disabilities: A comparison of linear, quadratic, and piecewise linear mixed effects models. *Journal of school psychology*, *53*(2), 105-120.

Kroesbergen, E. H., & Van Luit, J. E. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial and special education*, *24*(2), 97-114.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, *76*(4), 604-620.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in?. *Annals of Statistics, 27*(2), 415-438.

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities*, *41*(5), 451-459.

Lee, D. S. (2008). Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, *142*(2), 675-697.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281-355.

Lee, D. S., Moretti, E., & Butler, M. J. (2004). Do voters affect or elect policies? Evidence from the US House. *The Quarterly Journal of Economics, 119*(3), 807-859.

Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L., & MCSorley, E. O. (2005). Beyond the t-test: statistical equivalence testing. *Analytical Chemistry*, 77, 221A-226A.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*(1), 159–208.

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

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educational Researcher*, *45*(1), 18-35.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*(4), 306-321.

Moss, B. G., Yeaton, W. H., & LIoyd, J. E. (2014). Evaluating the effectiveness of developmental mathematics by embedding a randomized experiment within a regression discontinuity design. *Educational Evaluation and Policy Analysis*, *36*(2), 170-185.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: US Department of Education.

Nelson, N. J., Fien, H., Doabler, C. T., & Clarke, B. (2016). Considerations for realizing the promise of educational gaming technology. *Teaching Exceptional Children*, *48*(6), 293-300. https://doi.org/10.1080/19345747.2015.1119229

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional children*, *71*(2), 137-148.

Pettersson-Lidbom, P. (2008). Do parties matter for economic outcomes? A regression discontinuity approach. *Journal of the European Economic Association, 6*(5), 1037-1056.

R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www. R-project.org/

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

Reardon, S. F., & Stuart, E. A. (2017). Editor's introduction: Theme issues on variation in treatment effects. *Journal of Research on Educational Effectiveness*. *10*(4), 671-674.

Riecken, H. W., Boruch, R., Campbell, D. T., Caplan, N., Glenman, T. K., Pratt, J. W., et al. (1974). *Social Experimentation: A method for planning and evaluating social intervention*. New York: Academic Press.

Sacks, J., & Ylvisaker, D. (1978). Linear estimation for approximately linear models. *The Annals of Statistics*, *6*(5), 1122-1137.

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.

Schochet, P. Z. (2009). Statistical power for regression discontinuity design in educational valuations. *Journal of Educational and Behavioral Statistics*, *34*, 238-266.

Schochet, P. Z., Cook, T. D., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for Regression Discontinuity Designs*. What Works Clearinghouse. Retrieved from http://files.eric.ed.gov/fulltext/ED510742.pdf.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and non-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological methods*, *16*(2), 179-191.

Skovron, C., & Titiunik, R. (2015). *A Practical guide to regression discontinuity designs in political science*. Working Paper, University of Michigan.

Smith, W. C. (2014). Estimating unbiased treatment effects in education using a regression discontinuity design. *Practical Assessment, Research & Evaluation*, *19*(9). Available online: http://pareonline.net/getvn.asp?v=19&n=9.

Steiner, P. M. & Wong. V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation Review*, 1-34. DOI: 10.1177/0193841X18773807

Swanson, H. L., Orosco, M. J., & Lussier, C. M. (2014). The effects of mathematics strategy instruction for children with serious problem-solving difficulties. *Exceptional Children*, *80*(2), 149-168.

Tang, Y., & Cook, T. D. (2014). *Statistical power for the comparative regression discontinuity design with a pretest no-treatment control function: Theory and evidence from the National Head Start Impact Study*. Unpublished paper.

Tang, Y., Cook, T. D., Kisbu-Sakarya, Y., Hock, H., & Chiang, H. (2017). The comparative regression discontinuity (CRD) design: An overview and demonstration of its performance relative to basic RD and the randomized experiment. In M. D. Cattaneo & J. Escanciano (Eds.), *Regression Discontinuity Designs (Advances in Econometrics, Volume 38*, pp.237 – 279*)* Bingley, UK: Emerald Publishing Limited.

Thistlethwaite, D, & Campbell, D. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309-317.

Toll, S. W., & Van Luit, J. E. (2013). Accelerating the early numeracy development of kindergartners with limited working memory skills through remedial education. *Research in Developmental Disabilities*, *34*(2), 745-755.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). *Test of word reading efficiency (TOWRE)*. Austin, TX: Pro-Ed.

Trochim, W. M. K., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials*, *13*(3), 190-212.

Uppal, Y. (2009). The disadvantaged incumbents: estimating incumbency effects in Indian state legislatures. *Public Choice*, *138*(1-2), 9-27.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *Review of Economics and statistics*, *88*(1), 171-177.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, *99*(1), 179-215.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review*, *43*(4), 1249-1287.

Van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I. *Journal of Econometrics*, *142*(2), 731-756.

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: A missing link to creating optimally effective communication interventions. *Mental retardation and developmental disabilities research reviews*, *13*(1), 70-77.

What Works Clearinghouse (2017). Standards Handbook (Version 4.0). Washington, DC: US Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Whitley, E., & Ball, J. (2002). Statistics review 6: Nonparametric methods. *Critical Care, 6*(6). Available online: http://ccforum.biomedcentral.com/articles/10.1186/cc1820.

Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within study comparison. *Journal of Policy Analysis and Management*, *32*(4), 853-877.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of policy Analysis and Management*, *27*(1), 122-154.

Zvoch, K., Yoon, H., & Cook, T. D. (2016). *Application of a hybrid regression discontinuity design to examine the generality of program effects*. Paper accepted for presentation at the Annual Meeting of the Association for Public Policy Analysis and Management, Washington, DC.