Asia Pacific Education Review (2024) 25:629–645
https://doi.org/10.1007/s12564-024-09956-3

Regression discontinuity designs in education: a practitioner's guide

Youmi Suk[1]

Received: 19 June 2023 / Revised: 1 March 2024 / Accepted: 4 March 2024 / Published online: 10 April 2024

Abstract
Regression discontinuity (RD) designs have gained significant popularity as a quasi-experimental device for evaluating education programs and policies. In this paper, we present a comprehensive review of RD designs, focusing on the continuity-based framework, the most widely adopted RD framework. We first review the fundamental aspects of RD designs, drawing on potential outcomes and causal graphs. We then discuss the validity threats in RD designs, including manipulation, discreteness of the running variable, statistical power, and generalizability. Additionally, we provide an overview of the existing extensions to RD designs. To exemplify the application of RD methods, we analyze the effect of New Jersey's pre-kindergarten program on children's vocabulary test scores, using an educational dataset. Finally, we offer practical guidelines in the conclusion to promote the appropriate use of RD methods in educational research.

Keywords Regression discontinuity · Quasi-experimental designs · Nonexperimental methods · Causal inference · Program evaluation · Pre-kindergarten programs

Introduction

A regression discontinuity (RD) design has emerged as a prominent quasi-experimental design since its original conception by Thistlethwaite and Campbell (1960), and it is increasingly being used to evaluate programs or policies in education and the social sciences. Over the past two decades, researchers have devoted significant efforts to advancing the methodology and empirical application of RD designs. Several reviews have been conducted on these evolving designs, including works by Cook (2008), Imbens and Lemieux (2008), Lee and Lemieux (2010), and Cattaneo and Titiunik (2022). However, these reviews primarily originate from fields outside of education, and there is limited research that thoroughly examines the fundamental and practical aspects of RD designs in the context of education, along with clear demonstrations. The main goal of this paper is to provide a comprehensive review of both traditional RD designs and the latest developments that are particularly relevant in education settings. Additionally, the paper aims to illustrate key aspects of these designs using a real educational dataset and provide practical guidelines for implementing RD designs.

In educational programs, eligibility and enrollment policies often dictate the treatment assignment of students or children, typically based on factors such as age, abilities, or special needs. For example, state pre-kindergarten (pre-K) programs determine enrollment based on a child's date of birth. Children with birthdays on or after a specific date become eligible for enrollment in the pre-K programs, whereas those with birthdays before it do not qualify. This situation necessitates the use of RD designs, where the treatment assignment variable (also referred to as the running variable)—in this case, the birth date—completely determines the treatment status. Such an RD design is regarded as a quasi-experimental design that closely approximates a randomized experiment, given the known treatment assignment mechanism. That is, in the RD design, the variation in treatment assignment is as good as random near the cutoff when study units typically cannot control the running variable precisely near the cutoff (Lee and Lemieux, 2010). Furthermore, in an RD setting where treatment status is determined by the running variable, it is not feasible to employ a matching design, another popular quasi-experimental design in education. In the matching design, treated units and control units are matched based on the similarity of measured covariates in observed data in order to identify and estimate

✉ Youmi Suk
ysuk@tc.columbia.edu

[1] Department of Human Development, Teachers College Columbia University, Grace Dodge Hall 552, 525 West 120th Street, New York, NY 10027, USA

the average treatment effect (ATE) (Steiner and Cook, 2013). However, in the RD setting, there is a lack of overlap between treated and control units in terms of the running variable, making the use of matching strategies infeasible. Consequently, RD designs utilize distinct identification and estimation strategies compared to matching and other quasi-experimental designs.

Specifically, nonparametric identification of RD designs is limited to the ATE *at* or *around* the cutoff, as compatibility between the treated and control groups is achieved only in the close vicinity of the cutoff score. For the identification of the ATE at the cutoff in RD designs, two main frameworks are available: the continuity-based framework and the local randomization framework. The continuity-based framework, established by Hahn et al. (2001), provides valid counterfactuals by assuming that the conditional expectations of potential outcomes, given the running variable, are continuous at the cutoff. This assumption provides the fundamental requirement for nonparametric identification of the ATE at the cutoff; see Sect. "Basics of RD Designs" for details. On the other hand, the local randomization framework was initially motivated by the work of Lee (2008), which captures the original ideas in the seminal article by Thistlethwaite and Campbell (1960) and interprets RD designs heuristically as if they were randomly assigned in a small neighborhood of the cutoff. It was subsequently formalized by Cattaneo et al. (2015). The local randomization framework relies on a stronger assumption than the continuity assumption of the first framework, but it provides justification for employing estimation and inference methods from the analysis of experiments literature, such as Fisherian or Neyman approaches (Cattaneo et al., 2019a). In this paper, we primarily focus on the continuity-based framework due to its wide adoption and longer history. For those interested in the local randomization framework, refer to Cattaneo et al. (2015) and Cattaneo et al. (2017).

The RD literature has evolved in recent decades, with departures from the traditional RD design. Researchers have explored diverse issues, such as incorporating multiple cutoff values or running variables, utilizing coarse measurements of the running variable, handling multisite/multilevel data, and investigating different parameters of interest like regression kink designs; see Sect. "Extensions" for more details. In addition to methodological advancements, there has been a substantial increase in the practical application of RD methods. Recent studies employing RD designs in education have investigated a broad range of educational issues. For instance, these studies have focused on selective public schools, subsidized loan programs, algebra courses, reclassification of English language learners, test-based retention or remediation, school turnaround programs, supplemental reading or literacy programs, cash transfer programs, and testing accommodations. See Table 1 for a list of the recent publications.[1]

Table 1 Recent publications on regression discontinuity (RD) in education

Publication	Setting
Angrist and Rokkanen (2015)	Selective public schools
Bergolo and Galván (2018)	Cash transfer programs
Brunner et al. (2023)	Selective public schools
Carlson and Knowles (2016)	English language learner reclassification
Coyne et al. (2018)	Reading programs
Figlio et al. (2018)	Literacy programs
Figlio and Özek (2023)	Test-based remediation
Heissel and Ladd (2018)	School turnaround programs
Lee and Soland (2022)	English language learner reclassification
McEachin et al. (2020)	Algebra courses
Melguizo et al. (2015)	Subsidized loan programs
Nomi and Raudenbush (2016)	Algebra courses
Schwerdt et al. (2017)	Test-based retention
Suk et al. (2022)	Testing accommodations

The remainder of the paper is organized as follows. Section "Basics of RD Designs" offers a concise overview of the fundamentals of RD designs within the continuity-based framework, drawing on potential outcomes and causal graphs. Section "Threats to Validity of RD Designs" discusses the potential threats to the validity of RD designs and addresses concerns on the underlying assumptions. Section "Extensions" explores recent advancements and extensions in RD designs. Section "Educational Example: New Jersey's Pre-K Programs" presents an analysis of our empirical example concerning New Jersey's pre-K program. Conclusions with practical guidelines are provided in Sect. "Conclusions".

Basics of RD designs

RD designs come in two main types depending on whether study units comply with the assigned treatment status. In cases of perfect compliance, it is known as a *sharp* RD design, while with imperfect compliance, it is referred to as a *fuzzy* RD design. In the following section, we review the details of causal estimands, assumptions, and estimation methods for both of these types.

[1] For a list of other RD applications in education, refer to Table 5 in Lee and Lemieux (2010) and Secti. 4.2 of Villamizar-Villegas et al. (2021).

Notation

We use the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974) to define treatment effects. Let $A_i \in \{0, 1\}$ be a binary treatment variable where $A_i = 1$ indicates that child i was assigned to (or eligible for) the pre-K treatment and $A_i = 0$ indicates the control condition. In a classic RD design, treatment assignment is based on a continuous running variable X_i and a cutoff score x_c such that $A_i = 1$ if $X_i \geq x_c$ and $A_i = 0$ if $X_i < x_c$. Let $T_i \in \{0, 1\}$ denote the treatment received where $T_i = 1$ if child i actually received the pre-K program and $T_i = 0$ if the child did not receive the program. Note that when full compliance is achieved with respect to the assignment/eligibility rule, the assignment status and treatment received status are identical, i.e., $A_i = T_i$.

$Y_i(1)$ represents the potential treatment outcome if child i were to receive a pre-K program, and $Y_i(0)$ represents the potential control outcome for the same child but under the control condition. For every child, the observed outcome is linked to the potential outcomes as follows: $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. The equality implies the stable unit treatment value assumption (SUTVA; Rubin , 1986), that is, (1) a child's potential outcomes are independent of other children's treatment assignment and (2) there are no different versions of the treatment. Finally, let **W** be a set of observed pre-treatment covariates and **U** be unobserved covariates.

Sharp RD designs

In this subsection, we review the standard, sharp RD design with a continuous running variable. Let's assume that our running variable is a child's birth date, which is measured continuously in days. In this case, children with birthdays after or on the cutoff, denoted as $X_i \geq x_c$, are assigned to the pre-K program, but those with birthdays before the cutoff are not assigned and must wait another year. When there is full compliance with the assigned pre-K status, this design is called a sharp RD design. Under this design, the causal estimand of interest is the ATE at the cutoff, which is the average linear contrast of potential outcomes between the treated and control groups at the cutoff value of the running variable and is defined as τ_{SRD}:

$$\tau_{SRD} = E[Y_i(1) - Y_i(0) \mid X_i = x_c]. \quad (1)$$

In our example, the ATE at the cutoff represents the average effect of the pre-K program for children scoring at the eligibility cutoff. In sharp RD designs, the probability of receiving treatment abruptly changes from one to zero when the running variable X_i crosses the cutoff x_c. Since A_i is a known deterministic function of X_i, we achieve conditional unconfoundedness, meaning that $Y_i(1), Y_i(0) \perp A_i | X_i$. However, there is no common support between treatment and control units on X_i in the sharp RD design. This means that units scoring above or at the cutoff have a probability of 1 for receiving treatment ($Pr(A_i = 1 | X_i \geq x_c) = 1$), whereas those scoring below the cutoff have a probability of 0 ($Pr(A_i = 1 | X_i < x_c) = 0$). Therefore, due to the violation of the positivity assumption (i.e., $0 < Pr(A_i = 1 | X_i) < 1$), the treatment effect for the entire population cannot be non-parametrically identified. Nevertheless, the treatment effect for the subpopulation at the cutoff can still be non-parametrically identified if the potential outcomes satisfy the local continuity assumption at the limiting cutoff as follows:

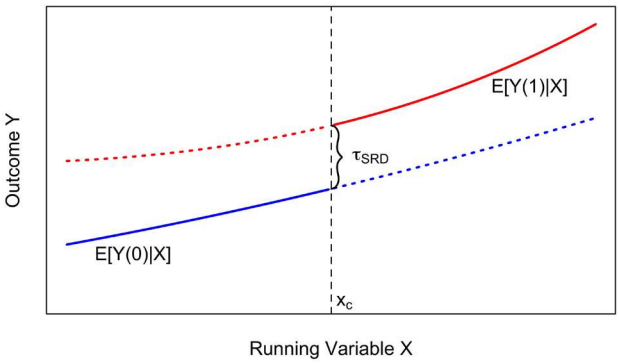


Fig. 1 Visual representation of a regression discontinuity (RD) design

(A1) Local Continuity of Potential Outcomes:

$$\lim_{x \uparrow x_c} E(Y_i(1) \mid X_i = x) = \lim_{x \downarrow x_c} E(Y_i(1) \mid X_i = x),$$
$$\lim_{x \uparrow x_c} E(Y_i(0) \mid X_i = x) = \lim_{x \downarrow x_c} E(Y_i(0) \mid X_i = x).$$

This assumption means that the average potential treatment and control outcomes just below the cutoff are equal to the respective average potential outcomes just above the cutoff. This assumption allows us to establish valid counterfactuals for children near the cutoff (Hahn et al., 2001; Imbens and Lemieux, 2008). Under Assumption [A1], the ATE at the cutoff, i.e., τ_{SRD}, is identified as follows:

$$E[Y_i(1) - Y_i(0) \mid X_i = x_c] = \lim_{x \downarrow x_c} E(Y_i \mid X_i = x) - \lim_{x \uparrow x_c} E(Y_i \mid X_i = x).$$

In Fig. 1, we illustrate an RD design with (unknown) potential treatment and control outcome functions represented by red and blue lines, respectively. The figure highlights that potential outcomes are continuous at the cutoff, with no overlap between treatment and control units in terms of the running variable. Solid lines can be estimated using

regression smoothers based on observed data, while dashed lines cannot be estimated due to the unavailability of data.

Before formally estimating the RD effect, a visual inspection is essential. Researchers create an RD plot by plotting the relationship between the running variable and the outcome of interest using a scatterplot and smoothing regression lines. This RD plot is similar to Fig. 1, but it is based on observed data and employs empirical regression smoothers (see an example in Figs. 4 in Sect. Educational example: New Jersey's pre-K programs). Plotting such RD plots plays a crucial role in RD empirical analysis and should precede the formal estimation of the RD treatment effect.

To formally estimate τ_{SRD}, researchers can employ various approaches, including parametric, semi-parametric, or nonparametric methods (e.g., Lee and Lemieux 2010; Imbens and Lemieux 2008; Schochet et al. 2010). First, the parametric approach fits a regression model that regresses the outcome on assignment status A_i and centered running variable $(X_i - x_c)$, as follows:

$$Y_i = \beta_0 + \beta_1 A_i + f(X_i - x_c) + \epsilon_i. \tag{2}$$

Here, the term β_0 represents the intercept of the control group, and $\beta_1 = \tau_{SRD}$ represents the ATE at the cutoff; $f(\cdot)$ represents a functional form of the running variable, and ϵ_i represents the random error. Typically, the regression slopes of the running variable differ between the left and right sides of the cutoff value by including an interaction term between A and X, as: $Y_i = \beta_0 + \beta_1 A_i + \beta_2(X_i - x_c) + \beta_3 A_i(X_i - x_c) + \epsilon_i$ (Lee and Lemieux, 2010). Additional higher-order terms (e.g., quadratic, cubic terms) can be incorporated into the regression model. The choice of the polynomial functional form can be based on the statistical significance of higher-order terms or model-fit criteria, such as the F-test statistic and Akaike Information Criteria (AIC). However, a limitation of using parametric regression is that it provides global estimates of the regression function across the entire range of X, rather than focusing on the subpopulation at the cutoff (Lee and Lemieux, 2010).[2]

Alternatively, nonparametric approaches, such as local polynomial regression, can be utilized and have become more widely used for RD estimation. In local polynomial regression, researchers need to determine the kernel function, bandwidth, and the inclusion of higher-order terms (Lee and Lemieux, 2010; Imbens and Lemieux, 2008). For example, a local linear regression model with a rectangular/uniform kernel (and different slopes) can be written as:

$$\begin{aligned} Y_i &= \beta_0' + \beta_1' A_i + \beta_2'(X_i - x_c) + \beta_3' A_i(X_i - x_c) + \epsilon_i', \\ w_i &= \begin{cases} 1, \text{ if } |(X_i - x_c)|/h < 1 \\ 0, \text{ if } |(X_i - x_c)|/h \geq 1 \end{cases} \end{aligned} \tag{3}$$

Here, the bandwidth h controls the width of the neighborhood around the cutoff and represents half of the window width. The weight $w_i = 1$ if observation i lies within the window and $w_i = 0$ if it is outside the window. This means that observations within the window have equal weight and observations outside the window are excluded from RD analysis. While other kernels (e.g., triangular or Epanechnikov) can also be used, the choice of kernel function usually has minimal impact. However, choosing bandwidth h is crucial and involves finding an optimal balance between precision and bias. A larger bandwidth produces more precise estimates because more observations are available for estimating the regression, but it introduces more smoothing bias to the local polynomial approximation. Furthermore, the choice of bandwidth affects the selection of higher-order terms as smaller bandwidths require lower higher-order terms (Lee and Lemieux, 2010; Cattaneo et al., 2019b).

Two main procedures are commonly used to select bandwidths. The first procedure involves characterizing the optimal bandwidth in terms of the unknown functionals (e.g., mean, variance) of the data distribution. These functionals can then be estimated using data-driven methods, and plugged into the optimal bandwidth function (Imbens and Kalyanaraman, 2012; Calonico et al., 2014, 2019). An increasingly popular method in the first procedure is to find the value of h by minimizing the mean square error (MSE) of the RD effect estimator $\hat{\tau}_{SRD}$, given a choice of polynomial order and kernel function (Cattaneo et al., 2019b). The second approach to choosing bandwidths is based on a cross-validation procedure as demonstrated by Ludwig and Miller (2007). This approach determines the optimal bandwidth by selecting the value of h that minimizes the MSE between the predicted and observed value of Y. Currently, the first procedure has become more popular because choosing a bandwidth that is optimal for estimating τ_{SRD} is more relevant in RD settings (Imbens and Kalyanaraman, 2012; Cattaneo et al., 2019b).

Regarding baseline covariates $\mathbf{W}_i$, it is not necessary to include them in RD analysis to obtain consistent estimates of the RD effect. However, the key advantage of incorporating baseline covariates into regressions is an efficiency gain. In other words, it allows us to improve the precision of the estimated effect if baseline covariates are correlated with the outcome variable. Additionally, when the window or bandwidth size is wide and we include observations that are farther away from the cutoff in RD analysis, using additional covariates can potentially help mitigate bias arising from these additional observations (Imbens and Lemieux,

[2] When the functional form of the regression model is uncertain, it is recommended to adopt an overfitting strategy by including more polynomial and interaction terms than strictly necessary (Shadish et al., 2002).

2008; Calonico et al., 2019). For more information on the estimation and inference in sharp RD designs, refer to Lee and Lemieux (2010), Cattaneo et al. (2019b), and Cattaneo and Titiunik (2022). Additionally, Sect. Educational example: New Jersey's pre-K programs of this paper provides a detailed demonstration.

Fuzzy RD designs

In practice, it is often observed that study administrators deviate from the assignment rules, or participants fail to comply with their assigned treatment status. For example, children who are eligible for a pre-K program based on their birth dates may not participate, and ineligible children might actually participate due to specific rules or exemptions. When such non-compliance occurs, the probability of receiving the treatment is less than one but greater than zero, i.e., $0 < \lim_{x \downarrow x_c} Pr(T_i = 1 \mid X_i = x) - \lim_{x \uparrow x_c} Pr(T_i = 1 \mid X_i = x) < 1$, and we have what is known as a *fuzzy* RD design.

In this design, two causal estimands are of interest: the intent-to-treat (ITT) effect and local average treatment effect (LATE), both at the cutoff score. The ITT effect at the cutoff is defined as Eq. (1). It is identified and estimated using the same approach as the ATE in the sharp RD design discussed in Sect. Sharp RD designs, where the pre-K assignment/eligibility status, A_i, serves as the "treatment" indicator. To define the LATE at the cutoff, we now use potential outcomes notations for treatment receipt. Let $T_i(1)$ denote a child's potential pre-K receipt if they were eligible ($A_i = 1$), and let $T_i(0)$ denote their potential non-receipt if they were ineligible ($A_i = 0$). We assume $T_i = A_i T_i(1) + (1 - A_i) T_i(0)$. Because both $T_i(0)$ and $T_i(1)$ are binary indicators, there are four possible values for the pair of potential responses to treatment assignment. The first group, referred to as *compliers* includes units who always comply with their assignment (i.e., $T_i(0) = 0, T_i(1) = 1$). In this study, compliers mean students who would receive the pre-K program if eligible and would not receive it if ineligible. All other units are classified as noncompliers, but they can be categorized into three distinct types: *never-takers*, *always-takers*, and *defiers*. Never-takers are units who never take the treatment, regardless of their assignment (i.e., $T_i(0) = 0, T_i(1) = 0$), whereas always-takers are those who would always take the treatment, regardless of their assignment (i.e., $T_i(0) = 1, T_i(1) = 1$). Finally, defiers are those who would act contrary to their assignment (i.e., $T_i(0) = 1, T_i(1) = 0$) (Imbens and Rubin, 2015).

Under the fuzzy RD design, the LATE at the cutoff, denoted as τ_{FRD}, is the ATE at the cutoff for the subpopulation of compliers, which is formally defined as follows:

$$\tau_{FRD} = E[Y_i(1) - Y_i(0) \mid X_i = x_c, T_i(1) = 1, T_i(0) = 0] \quad (4)$$

In our setting, τ_{FRD} represents the average effect of receiving the pre-K program for students who comply with the treatment assigned status at the cutoff. To identity the LATE at the cutoff, the fuzzy RD design make two additional assumptions:

(A2) Local Monotonicity:

$$\lim_{x \uparrow x_c} Pr(T_i(1) < T_i(0) \mid X_i = x) = \lim_{x \downarrow x_c} Pr(T_i(1) < T_i(0) \mid X_i = x) = 0$$

(A3) Local Exclusion Restriction:
$Pr(Y_i(1, t) \neq Y_i(0, t) \mid X_i = x_c) = 0$ for each $t = 0, 1$, and where the potential outcomes $Y_i(a, t)$ are now functions of both the treatment assigned/eligible status (a) and the treatment received status (t).

The local monotonicity assumption ensures the absence of defiers at the cutoff, which, in our empirical example, refers to children who would receive the pre-K program if ineligible but would not receive it if eligible. The local exclusion restriction assumption states that potential outcomes depend solely on treatment receipt (T_i) and are unaffected by treatment assignment (A_i) at the cutoff; that is, under this assumption, i.e., $Y_i(a, t) = Y_i(t)$. Assumptions [A1]-[A3] allow us to identify the LATE at the cutoff as follows:

$$E[Y_i(1) - Y_i(0) \mid X_i = x_c, T_i(1) = 1, T_i(0) = 0] = \frac{\lim_{x \downarrow x_c} E(Y_i \mid X_i = x) - \lim_{x \uparrow x_c} E(Y_i \mid X_i = x)}{\lim_{x \downarrow x_c} E(T_i \mid X_i = x) - \lim_{x \uparrow x_c} E(T_i \mid X_i = x)}.$$

To estimate τ_{FRD} in fuzzy RD designs, researchers have the option to use either parametric regression with polynomial and interaction terms or nonparametric regression (Imbens and Lemieux, 2008; Lee and Lemieux, 2010). In the parametric approach, instrumental variable regression or two-stage least squares (TSLS) regression can be employed. However, similar to sharp RD designs, a limitation of this parametric approach is that it relies on all values of X without exclusively using the subpopulation near the cutoff. On the other hand, nonparametric regression focuses on a small neighborhood around the cutoff point, where local polynomial regression is commonly used to estimate the numerator and denominator of the ratio in τ_{FRD} (Lee and Lemieux, 2010). Like sharp RD designs, researchers need to select the kernel function, bandwidth, and the inclusion of higher-order terms for both the treatment regression and outcome regression. In practical applications, it is desirable to use the same bandwidth for both the numerator and denominator (assuming the same kernel function and polynomial order). This enhances transparency and simplifies computation in RD analysis, because researches can clarify which observations are included in their calculations. Similar to sharp RD designs, the bandwidth can be chosen using an optimal

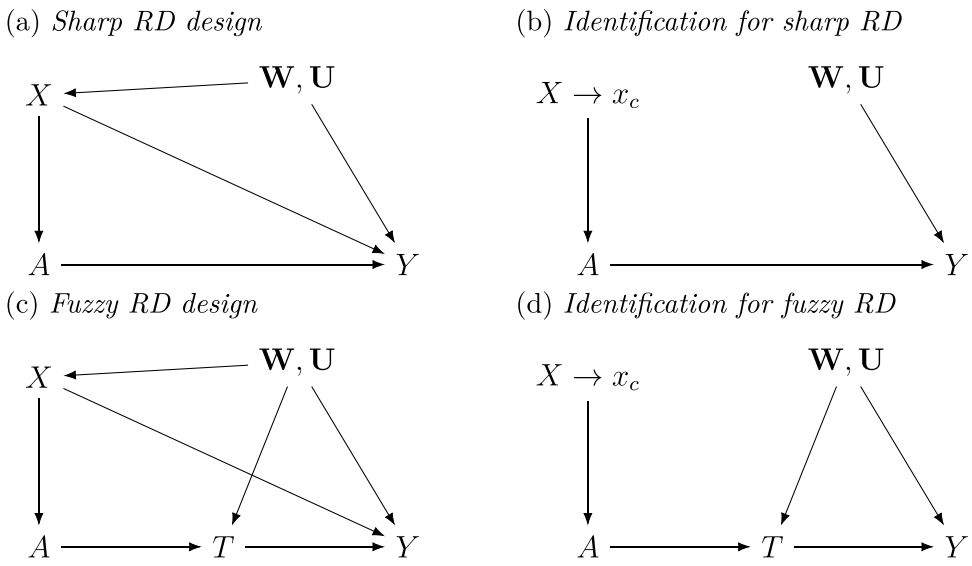


Fig. 2 Causal graphs and causal graphical identification for evaluating a pre-K program based on sharp and fuzzy RD designs. X represents the running variable, which is a child's birth date. A represents the pre-K assigned status. T represents the pre-K received status. Y represents the outcome, which is vocabulary test scores. $\mathbf{W}$ represents measured covariates, and $\mathbf{U}$ represents unmeasured covariates

data-driven approach that aims to minimize the MSE of the fuzzy RD effect estimator. Alternatively, researchers can determine the optimal bandwidth through a cross-validation procedure that minimizes the MSE between the predicted and observed outcomes.[3] Lastly, incorporating covariates $\mathbf{W}_i$ into the analysis can enhance the efficiency of the estimator and may help mitigate potential covariate imbalance arising from the inclusion of observations that are farther away from the cutoff within the window. For more information on the estimation and inference in fuzzy RD designs, refer to Imbens and Lemieux (2008), Lee and Lemieux (2010), and Cattaneo et al. (2019a). An example of implementation is discussed in Sect. Educational example: New Jersey's pre-K programs.

A graphical perspective on RD designs

Causal graphs, known as directed acyclic graphs (DAGs), provide a useful framework to describe the causal relationships between variables and offer a formal yet intuitive discussion of causal identification in both sharp and fuzzy RD designs (Pearl, 2009; Steiner et al., 2017). We use the data-generating models underlying RD designs as represented by causal graphs in Fig. 2. In the figure, the pre-K assignment/eligibility status (A) is exclusively determined by the running variable, a child's birth date (X). In Fig. 2a for the sharp RD design, the running variable X affects the eligibility status A and the outcome Y, and thus, it confounds the causal relationship between A and Y. Observed and unobserved covariate sets ($\mathbf{W}, \mathbf{U}$) affect the running variable X and the outcome Y, which introduces confounding in the causal relationship between A and Y via X. Although conditioning on X blocks the confounding backdoor paths between A and Y,[4] the ATE of A on Y for the overall population remains unidentified due to the violation of the positivity assumption, i.e., $0 < Pr(A = 1|X) < 1$. That is, we lack the overlap of the running variable, meaning that eligible and ineligible students are situated in non-overlapping regions of the running variable. Thus, we leverage the discontinuity at the cutoff, instead of matching methods, to identify the ATE of A on Y at the cutoff.

Figure 2b demonstrates the graphical identification for the sharp RD design at the limiting cutoff score, $X \to x_c$ (Steiner et al., 2017). In this scenario, the running variable (X) still determines pre-K eligibility (A), but it no longer directly affects the outcome (Y) nor is it influenced by the measured and unmeasured covariates ($\mathbf{W}$ and $\mathbf{U}$). Consequently, in the proximity of the cutoff score, pre-K eligibility becomes independent of $\mathbf{W}$ and $\mathbf{U}$, allowing for the identification of the ATE at the cutoff without any adjustments for covariates. Note that incorporating measured covariates $\mathbf{W}$ enhances the efficiency of the treatment effect by explaining the variance of Y.

[3] Note that in fuzzy RD designs, it is recommended to select the bandwidth based on the outcome regression and then use the same bandwidth for the treatment regression. This recommendation is based on the observation that the treatment regression typically requires a wider bandwidth, as it is expected to exhibit a very flat relationship.

[4] The backdoor criterion in causal graphs (Pearl, 1995) involves identifying and adjusting for variables that lie on "backdoor paths" between the treatment and outcome variables. By conditioning on these variables, non-causal paths are blocked, and this blocking allows for unbiased estimation of causal effects in observational studies.

On the other hand, in Fig. 2c for a fuzzy RD design, the pre-K eligibility status (A) differs from the pre-K receipt status (T), which depends on the eligibility status (A) and covariates ($\mathbf{W}, \mathbf{U}$). Administrators may offer the pre-K program to ineligible children or withhold them from eligible children based on the values of $\mathbf{W}$ and $\mathbf{U}$. Measured covariates $\mathbf{W}$ may include child background variables like gender, race/ethnicity, or free lunch status. Unmeasured covariates $\mathbf{U}$ may include a child's developmental immaturity or prior academic performance. In the fuzzy RD design, as the covariate sets $\mathbf{W}$ and $\mathbf{U}$ influence T and Y, they introduce confounding in both the ATE of A on Y and the ATE of T on Y for the overall population. Consequently, without covariate adjustments, these two ATEs remain unidentified. However, the presence of unmeasured covariates $\mathbf{U}$ renders matching methods infeasible. Instead, we leverage the discontinuity at the cutoff.

Figure 2d illustrates the graphical identification for the fuzzy RD design at the limiting cutoff score, $X \to x_c$. Similar to the sharp RD design, the running variable (X) solely determines pre-K eligibility (A) at the cutoff without any relationship with covariates $\mathbf{W}$ and $\mathbf{U}$. Thus, the ITT at the cutoff, representing the effect transmitted along $A \to Y$, can be identified by limiting $X \to x_c$, without covariate adjustments for $\mathbf{W}$ and $\mathbf{U}$. Moreover, using A as an instrument for pre-K receipt (T) enables the identification of LATE at the cutoff, i.e., the effect of $T \to Y$. It should be also noted that incorporating measured covariates $\mathbf{W}$ improves the efficiency of the treatment effect by accounting for the variance in T and Y.

Threats to validity of RD designs

This section discusses potential threats to the validity of RD designs, focusing on four main issues: manipulation or sorting around the cutoff, discreteness of the running variable, statistical power, and generalizability.

Manipulation or sorting around the cutoff

RD designs are appropriate and considered as good as randomized experiments when individuals cannot manipulate the running variable to precisely sort around the cutoff value. Manipulation refers to the systematic changes of values of the running variable for some units to influence treatment assignment (Schochet et al., 2010). Manipulation of the running variable can occur when the treatment has significant benefits or harms. For example, if the running variable for assigning a beneficial program is self-reported age with a publicly known cutoff value, one might see relatively more individuals with a reported age just below (or just above) the cutoff to participate (or not participate) in the program. This manipulation undermines the validity of the RD design and hinders the identification of the RD treatment effect, because units just below are no longer comparable to those just above. To assess whether the underlying assumption of individuals' inability to precisely manipulate the assignment variable is unwarranted, two types of tests are available; one examines the density continuity of the running variable, and the other examines the continuity of covariate distributions, both at the cutoff. One advantage of the former test is that it can be always conducted in an RD setting, while the latter test depends on the availability of data on these covariates. If either test yields significant results, it challenges the validity of the continuity assumption.

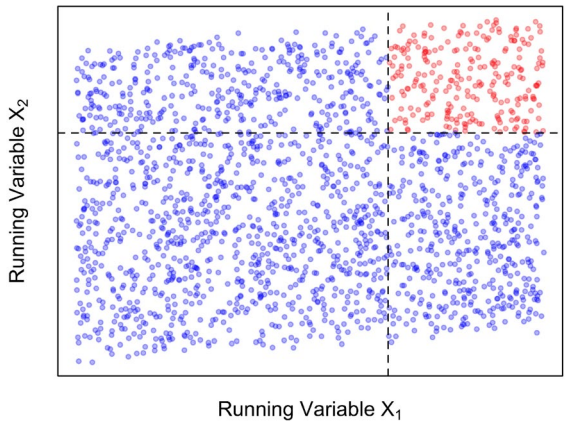
A direct and straightforward test is on examining whether the density of the running variable is continuous at the cutoff. Such tests include the McCrary's test based on a density function (McCrary, 2008), an empirical likelihood testing procedure (Otsu et al., 2013), and a local polynomial density estimator (Cattaneo et al., 2018, 2020). While a continuous density of the running variable at the cutoff is not by itself sufficient to confirm the validity of an RD design, a discontinuous density indicates endogenous sorting of units around the cutoff and should raise serious doubts about the appropriateness of the RD design (Cattaneo and Titiunik, 2022).

Another way to test the validity of the RD design is to examine whether baseline covariates are locally balanced or continuous on either side of the cutoff. If individuals cannot precisely manipulate the assignment variable in the RD design, the treatment assignment is locally randomized at the cutoff, and individuals in close proximity to the cutoff are expected to be comparable in terms of baseline covariates. That is, the baseline covariates are locally balanced on either side of the cutoff. Therefore, the distributions of observed baseline covariates should not change discontinuously at the cutoff (Imbens and Lemieux, 2008; Schochet et al., 2010; Lee and Lemieux, 2010). If the distribution of each covariate (conditional on the running variable) is discontinuous, it suggests non-random sorting of units into groups based on the given covariate. To test for non-random sorting based on baseline covariates, researchers can use estimation methods, such as local polynomial regression discussed in Sect. Sharp RD designs, with the response being each measured covariate instead of the outcome variable. However, it is not possible to test whether unmeasured covariates are continuously associated with the running variable at the cutoff. We provide an illustration of these falsification tests in Sect. Falsification tests.

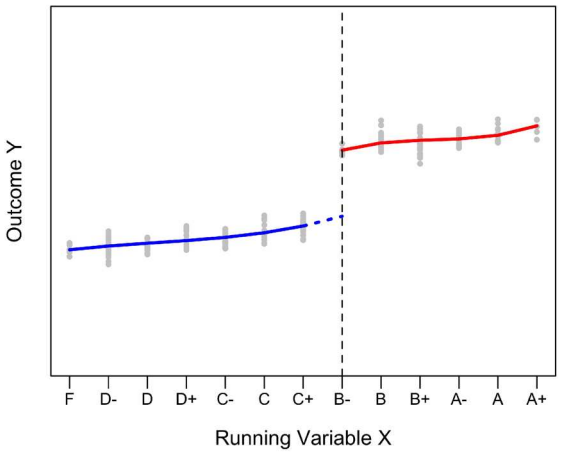
Discreteness of the running variable

In practice, the running variable is often measured on a discrete scale, represented by $X \in \{x_1, x_2, \ldots, x_{c-1}, x_c, \ldots x_K\}$ with K discrete values. For example, the running variable of

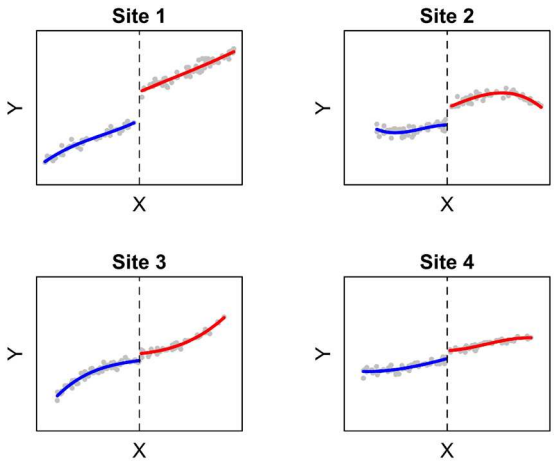
(a) *Multiple running variables*



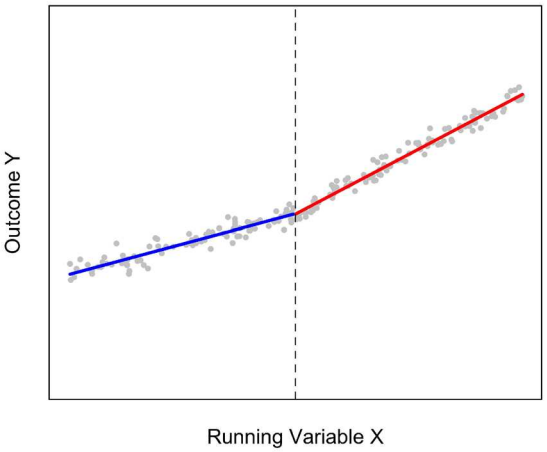
(b) *Ordinal running variable*



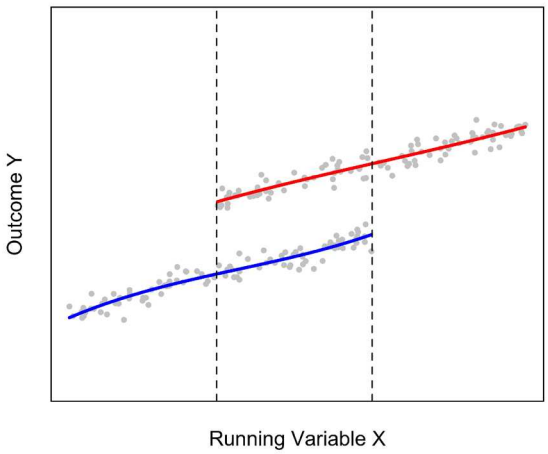
(c) *Multisite RD design*



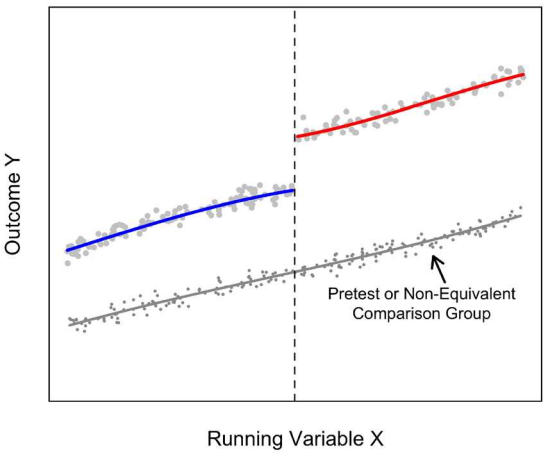
(d) *Regression kink design*



(e) *Combined RD and randomized experiment*



(f) *Comparative RD design*



◂**Fig. 3** Extensions in regression discontinuity (RD) designs. Dashed lines indicate the cutoff points. Red and blue colors are associated with observed treatment and control units, respectively. **a** RD design with multiple running variables, denoted as X_1 and X_2. **b** RD design with an ordinal running variable, such as the final letter grade for academic performance. **c** Multisite RD design, involving multiple sites (e.g., schools, hospitals). **d** Regression kink design, targeting the slope difference at the cutoff. **e** RD design combined with a randomized experiment. **f** Comparative RD design with a pretest or nonequivalent comparison group

a child's birth date would be measured based on intervals of 5 days rather than on a daily basis. Despite this discretization of the running variable, when certain conditions are met, researchers can still utilize the identification and estimation strategies discussed in Sect. Basics of RD designs. These conditions include (a) the accurate and precious (implicit) extrapolation from $X_i = x_{c-1}$ to $X_i = x_c$ and (b) a large number of unique values K (Cattaneo and Titiunik, 2022).

However, when the running variable has only a few distinct values, such as a child's birth date measured in months, the aforementioned identification and estimation strategies are not valid for analyzing RD designs. In such a case, Lee and Card (2008) propose using regressions to estimate the conditional expectation of the outcome variable at the cutoff point through extrapolation. This approach assumes that the parametric functional form on both sides of the cutoff score is correctly specified, which enables accurate extrapolation to the cutoff score. Additionally, the statistical uncertainty arising from the discreteness of the running variable should be considered by estimating cluster standard errors (Lee and Card, 2008). There are alternative approaches for RD designs with a discrete running variable, including one proposed by Dong (2014) that specifically addresses rounding errors in the running variable.

Low statistical power

In general, RD designs exhibit lower statistical power compared to randomized experiments with equal sample sizes due to greater sampling variance. Consequently, RD designs typically require much larger sample sizes to achieve the same level of statistical power (Goldberger, 1972; Shadish et al., 2002; Schochet, 2009). Statistical power depends on factors such as the significance level, effect size, sample size, and is also influenced by measurement error. Specifically, research by Goldberger (1972) revealed that for nonclustered designs, an RD design typically requires a sample size 2.75 times larger than a corresponding experiment to achieve the same level of statistical precision. For clustered designs, Schochet (2009) found that three to four times larger samples are usually required in RD designs compared to experimental clustered designs to attain the same level of precision. The reduced precision in RD designs arises due to the inherent correlation between treatment assignment and running variables included in the regression models, but this correlation is absent in randomized experiments (Schochet, 2009). However, when working with large-scale educational datasets, sample sizes are generally more than adequate to ensure sufficient power. Therefore, this potential limitation of RDD is often less problematic in practice compared to the other challenges we discuss.

Measurement error in data can also reduce statistical power in RD settings as in randomized experiments or other quasi-experimental designs. When the measurement error is not properly accounted for, it leads to an overestimation of power. Specifically, while the measurement error in the outcome is unlikely to introduce bias in the RD treatment effect, it increases the uncertainty of the effect estimate, making it more challenging to distinguish true effects from random variation or measurement error (Shadish et al., 2002). Therefore, in the presence of measurement outcome error, the minimum detectable effect (i.e., the smallest true effect size that can be detected) will increase, and thus, a larger sample size is typically required to achieve the same level of precision.

To account for potential power issues, researchers planning new experiments or surveys in RD designs can conduct power calculations and determine the required sample size at the design stage. Cattaneo et al. (2019) discuss power calculations and optimal sample size selection using local polynomial estimation and inference methods in RD designs, and Schochet (2009) and Bulus (2021) discuss power issues using parametric regression specifications for clustered RD designs.

Limited generalizability

As mentioned earlier, the RD treatment effect applies specifically to the subpopulation of individuals at or around the cutoff value, as there is no overlap of the running variable. While this allows for the identification of the treatment effect within this subpopulation, it does not provide information about the effect in other subpopulations or the entire population. Without strong assumptions justifying extrapolation to other subpopulations (such as homogeneity of the treatment effect or parametric modeling assumptions on the treatment effect), it is not possible to estimate treatment effects away from the cutoff or the overall ATE (Imbens and Lemieux, 2008).

To address the extrapolation or generalizability in RD designs, several approaches have been proposed. These include using external pre-treatment measures and parametric imputation methods (Mealli and Rampichini, 2012; Wing and Cook, 2013), incorporating pre-treatment covariates under local conditional ignorability (Angrist and Rokkanen,

2015), employing a local extrapolation method via marginal treatment effects (Dong and Lewbel, 2015), utilizing multiple measures of the running score with a factor model (Rokkanen, 2015), and employing multiple cutoffs under the constant bias assumption (Cattaneo et al., 2020).

Extensions

This section explores methodological advancements and extensions in RD designs that are particularly relevant in education settings. These extensions involve incorporating multiple running variables, utilizing an ordinal running variable, analyzing multisite/multilevel data, and introducing a novel parameter of interest from regression kink designs. We also briefly discuss variations that combine RD designs with other existing causal inference methods. Figure 3 visually illustrates each extension, and we provide a concise overview of their key features below.

RD designs with multiple running variables

Multiple running variables are commonly employed for assigning units to treatment conditions, especially when additional exclusion or inclusion criteria are present. Suppose students are assigned to a gifted program based on their scores in two tests, denoted as X_{1i} and X_{2i}, where the first test measures reading ability and the second test measures math ability, as depicted in Fig. 3a. If a student's scores in both reading and mathematics are at or above the specified cutoff scores x_{c_1} and x_{c_2}, respectively, they are assigned to the gifted program, i.e., $T_i = I(X_{1i} \geq x_{c_1})I(X_{2i} \geq x_{c_2})$, where $I(\cdot)$ is the indicator function. When using multiple running variables, there is an infinite collection of cutoff points where the treatment assignment sharply changes from zero to one, as shown in Fig. 3a. This motivates the use of a treatment effect curve, which incorporates infinitely many cutoff points, rather than focusing on a single-point treatment effect. Various approaches exist for conducting RD designs with multiple running variables, such as response surface RD analysis, which utilizes the multidimensional response surface, and frontier RD analysis, which estimates pairwise treatment effects using a subset of the data. For more details on multiple running variables, refer to Reardon and Robinson (2012) and Wong et al. (2013).

A special case of RD designs with multiple running variables is a *geographic RD design*, where latitude and longitude in coordinate systems serve as the running variables. In this design, units receive treatment if they are located within a specific geographic area, whereas they do not receive it in adjacent areas. Geographic RD designs are particularly useful for evaluating programs or policies that operate differently in cities or states located near borders. Standard RD estimation methods that include two running variables can be applied to geographic RD designs with minor adjustments. For further information on geographic RD, refer to Keele and Titiunik (2015).

RD designs with an ordinal running variable

In practice, the running variable in RD designs can often be measured on an ordinal scale. Examples of ordinal running variables include the final letter grade for academic performance (e.g., $A^+, A, A^-, B^+, \ldots, F$, as illustrated in Fig. 3b) and English proficiency levels of English language learners (Suk et al., 2022), bond ratings (Li et al., 2021), and inmate classification scores (Hjalmarsson, 2009). Using an ordinal running variable presents challenges due to the lack of a meaningful scale of distance. They also have limited observations in a small neighborhood below (or above) the cutoff, which requires extrapolation as depicted with a dashed line in Fig. 3b.

There are some approaches for conducting RD designs with ordinal running variables, notably discussed in Suk et al. (2022) and Li et al. (2021). Suk et al. (2022) employ a scale function to map the ordinal running variable onto a numeric scale and incorporate parametric modeling assumptions on the outcome (and treatment) for causal identification in RD settings. They also present sensitivity analyses to check the conclusions' robustness to different design factors, such as the choice of the scaling function, the choice of the cutoff value, and unobserved confounding due to model misspecification. On the other hand, Li et al. (2021) utilize propensity scores as a surrogate continuous running variable, and unlike Suk et al. (2022), this approach is under the local randomization framework; see Li et al. (2021) for more details.

Multisite RD designs

Multisite RD designs are often employed in education settings where the treatment or intervention is implemented within sites (e.g., schools). These designs introduce several additional considerations not encountered in non-clustered data, including the heterogeneity of the treatment effect across sites, the use of different cutoff values or running variables across sites, and the endogeneity of the study design, which is typically influenced by site sizes and the proportions of treated units. Figure 3c provides an illustration of multisite RD designs with the same running variable and the same cutoff across sites. The figure demonstrates that the RD treatment effects vary depending on the sites, with Site 1 showing the largest effect, and that the proportions of

treated units are not constant across sites, with Site 2 having the highest proportion.

When using multisite RD designs, researchers typically estimate site-specific RD effects and then combine these estimates to obtain a single, pooled RD treatment effect through meta-analysis, similar to multisite randomized trials. Multisite RD designs often aim to target cross-site heterogeneity of treatment and estimate the cross-site treatment effect variance using random effects models or fixed intercepts random coefficient models (Nomi and Raudenbush, 2016; Lee and Soland, 2022; Brunner et al., 2023; McEachin et al., 2020). However, methodological advancements for multisite RD designs have been progressing slowly, and to the best of our knowledge, practical guidelines for implementing multisite RD designs are not yet available.

Regression kink designs

Regression kink designs, originally introduced by Nielsen et al. (2010), are employed when a treatment is determined by a known assignment rule that alters the slope between the treatment and the running variable at a specific cutoff point (referred to as the kink point) (Card et al., 2015, 2017). As depicted in Fig. 3d, in the kink design, the expectation is that the outcome regression function will be continuous at all values of the running variable, but its slope will exhibit discontinuity at the cutoff point. Therefore, instead of focusing on a vertical gap at the cutoff as in conventional RD designs, the kink design examines whether the slope of the relationship between the outcome and the running variable shows a kink or discontinuity at the cutoff point. Any observed kink in the outcome, given the comparability of individuals on either side of the kink point, can be attributed to the treatment effect (Card et al., 2015). Estimation methods for kink treatment effects can be adapted from RD estimation methods, such as employing (local) polynomial regressions. However, the focus is on estimating the first derivatives of the regression functions rather than estimating a shift in the intercept. For more details of regression kink designs, refer to Card et al. (2017).

Other variations

There are several other research designs that combine RD designs with existing causal inference methods. On one hand, RD designs can be combined with randomized experiments by utilizing a cutoff interval between two cutoff values, as shown in Fig. 3e. Within this interval, units are randomly assigned to treatment conditions, whereas units above or below the interval are assigned to a single condition. This combination increases the statistical power for testing treatment effects and enhances external validity (Shadish et al., 2002).

On the other hand, RD designs can be combined with other quasi-experimental designs, such as matching or propensity score designs (e.g., Linden and Adams 2012), difference-in-differences (e.g., Grembi et al. 2016) (often referred to as difference-in-discontinuities designs), interrupted time series (e.g., Hausman and Rapson 2018) (often referred to as regression-discontinuity-in-time designs), and multiple control groups (Suk and Kim, 2023). Moreover, RD designs can incorporate an additional untreated outcome comparison function, like a pretest or a comparison group's posttest, as illustrated in Fig. 3f. Integrating this untreated comparison data to RD designs is often referred to as comparative RD designs (Wing and Cook, 2013a; Tang et al., 2017). These combined designs enhance RD designs in various ways, serving as sensitivity analyses, highlighting focused developments in specific settings (e.g., time-series data), or improving efficiency. The integration of regression discontinuity with other quasi-experimental designs is an emerging field, and there are many opportunities for further exploration.

Educational example: New Jersey's pre-K programs

Data and methods

State pre-K programs are educational initiatives that receive funding or oversight from the state and are often administrated by local school districts (Wong et al., 2007). They can produce short-term effects, such as improved academic performance during early school years, and may also have long-term effects like higher high school graduation rates (Campbell and Ramey, 1994; Wong et al., 2007). Wong et al. (2007) used RD designs to evaluate the effects of five-state pre-K programs on children's vocabulary, math, and print awareness skills, using a child's birth date as the running variable. The data collection focused on pre-K programs for 4-year-olds. In this paper, we used the data from one state, New Jersey, and evaluated the effect of New Jersey's Abbott Preschool Program, one of three state-funded pre-K initiatives in the state. The data in New Jersey were obtained through stratified random sampling, with stratification based on factors such as district enrollment, geographic location, and urban versus rural area.

For our data analysis, we used vocabulary test scores as the outcome variable Y_i. As mentioned earlier, pre-K eligibility A_i is a binary variable, where $A_i = 1$ denotes that a child was eligible for the state pre-K program, and $A_i = 0$ denotes that a child was ineligible for the program. The eligibility status was determined based on a child's birth date, our running variable X_i, with the cutoff date being October 1st. However, the assignment status A_i is not the

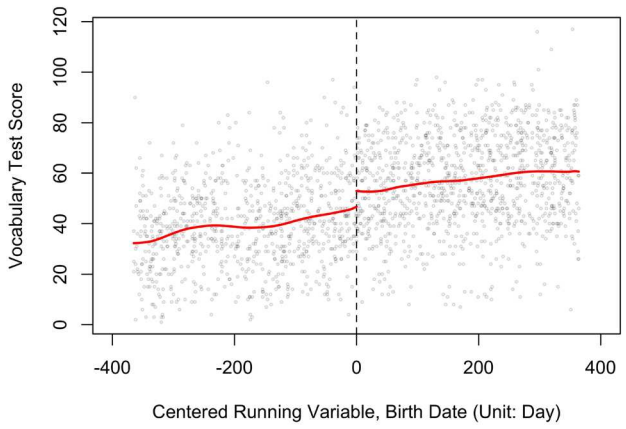


Fig. 4 Regression discontinuity plot for evaluating the average treatment effect of New Jersey's pre-K program in the vocabulary test

same as the receipt status T_i due to non-compliance. Specifically, the data exhibit two-sided non-compliance; some eligible students did not receive the pre-K program ($A_i = 1$ but $T_i = 0$), while some ineligible students received it ($A_i = 0$ but $T_i = 1$). In our analysis, we used a set of pre-treatment covariates $\mathbf{W}_i$, including gender, race/ethnicity, free lunch status, and test language type (English or Spanish), in the outcome and treatment regressions. We constrained our target sample to ±365 days from the cutoff date and excluded 9 cases with missing values in the outcome variable. As a result, our analytic sample consisted of 1,993 children, which accounted for 96.2% of the original sample in New Jersey.

Since there was non-compliance, we employed a fuzzy RD design and estimated the ITT and LATE of the state pre-K program at the cutoff. To estimate the ITT effect at the cutoff, we used local linear regression with a triangular kernel. The local linear regression model for the ITT at the cutoff with different slopes is written as follows:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 A_i + \beta_2 (X_i - x_c) + \beta_3 A_i (X_i - x_c) + \epsilon_i, \\ A_i &= \begin{cases} 1, \text{ if } X_i \geq x_c \\ 0, \text{ if } X_i < x_c, \end{cases}, \\ w_i &= \begin{cases} 1 - |(X_i - x_c)|/h, & \text{if } |(X_i - x_c)|/h < 1 \\ 0, & \text{if } |(X_i - x_c)|/h \geq 1 \end{cases} \end{aligned} \tag{5}$$

Here, the term $\beta_1 = \tau_{SRD}$ represents the ITT at the cutoff. The triangular kernel weight $w_i = 1 - |(X_i - x_c)|/h$ if observation i lies within the window and $w_i = 0$ if it is outside the window. This means that observations closer to the cutoff receives larger weights, while observations with $w_i = 0$ are excluded from RD analysis. We chose the optimal data-driven bandwidth that minimizes the MSE of the RD effect estimator $\hat{\tau}_{SRD}$ given our empirical data. Note again that the choice of kernel typically has little impact in practice when both the bandwidth and the polynomial order are fixed. We computed conventional confidence intervals based ordinary least squares (OLS) and robust (bias-corrected) confidence intervals. The robust confidence interval eliminates misspecification bias in the asymptotic approximation and adjusts the standard error to reflect extra variability from bias estimation. Unlike conventional intervals, this robust version is based on a valid standard normal distributional approximation (Calonico et al., 2014). We also used additional adjustment for measured covariates $\mathbf{W}$ potentially to improve efficiency.

To estimate the LATE at the cutoff, we employed local linear regression with kernel weights for both the treatment and the outcome. In this approach, we treated the model for T_i as the first-stage regression and the model for Y_i as the second-stage regression. This TSLS model for T_i and Y_i with triangular kernel weights can be written as follows:

$$T_i = \alpha_0 + \alpha_1 A_i + \alpha_2 (X_i - x_c) + \alpha_3 A_i (X_i - x_c) + \epsilon_i^t, \tag{6}$$

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 \hat{T}_i + \gamma_2 (X_i - x_c) + \gamma_3 A_i (X_i - x_c) + \epsilon_i^y, \\ A_i &= \begin{cases} 1, \text{ if } X_i \geq x_c \\ 0, \text{ if } X_i < x_c, \end{cases}, \\ w_i &= \begin{cases} 1 - |(X_i - x_c)|/h, & \text{if } |(X_i - x_c)|/h < 1 \\ 0, & \text{if } |(X_i - x_c)|/h \geq 1 \end{cases} \end{aligned} \tag{7}$$

In the first-stage regression (6), α_1 captures the discontinuity in the treatment probabilities between eligible and ineligible children at the cutoff. In the second-stage regression (7), we use the predicted value $\hat{T}_i$ as a regressor instead of A_i. This enables us to estimate $\gamma_1 = \tau_{FRD}$, representing the LATE at the cutoff value. Similar to the ITT estimation, we chose a MSE-optimal data-driven bandwidth, and we computed conventional confidence intervals and robust confidence intervals. We also did additional adjustments that include measured covariates in Eqs. (6) and (7) to improve efficiency.

Finally, we conducted falsification tests, as discussed in Sect. Manipulation or sorting around the cutoff, specifically focusing on manipulation or non-random sorting near the cutoff. For software, we used the R package *rdrobust* (Calonico et al., 2023) to conduct RD analysis, and we also used the R package *rddensity* (Cattaneo et al., 2023) to perform a manipulation test based on density discontinuity. R codes for our data analysis can be found at the first author's GitHub repository (https://github.com/youmisuk/RDDreview).

Results

ITT at the cutoff

Figure 4 presents a visual representation of the RD design for the ITT at the cutoff, i.e., the ATE at the cutoff, where the

Table 2 Intent-to-treat (ITT) effect of New Jersey's pre-K program at the cutoff

	Bandwidth	Estimate	Effect Size	95% CI	95% Robust CI
ITT (Optimal BW)	97.50	6.26	0.32	[0.47, 12.06]	[−0.04, 13.66]
Optimal BW w/ Covs	97.50	4.19	0.21	[−1.28, 9.65]	[−7.03, 9.52]
Half-BW	48.75	5.97	0.30	[−1.98, 13.93]	[−7.06, 14.64]

Effect sizes are calculated using the empirical sample standard deviation of the outcome

CI represents the confidence interval, *BW* represents the bandwidth, and *Covs* represents covariate adjustment

Table 3 Compliance for New Jersey's pre-K program by pre-K eligible status

Eligibility	Non-Received	Received	Total
Ineligible	852	19	871
Eligible	21	1,101	1122
Total	873	1,120	1993

x-axis represents the centered running variable with the cutoff set to zero. The figure includes a local linear regression line (red solid line) with an optimal, data-driven bandwidth of 97.5 days. Clearly, there is a discontinuity at the cutoff, and this potentially indicates the presence of a positive ATE at the cutoff.

Table 2 provides a summary of the ITT results for with and without additional covariate adjustment as well as different bandwidth choices. We report the effect size estimates by dividing the ITT estimates by the sample standard deviation of the outcome (19.83 points), and also report conventional OLS confidence intervals and robust (bias-corrected) confidence intervals. The ITT estimate using the MSE-optimal bandwidth of 97.5 days was 6.26 points, but it was not statistically significant with respect to the robust confidence interval. The corresponding effect size was 0.32, indicating a small effect based on Cohen's criterion.

We conducted sensitivity checks with the inclusion of covariates or the bandwidth size. Including covariates in RD analysis can enhance efficiency and potentially adjust covariate imbalance in the subsample within the window. Consequently, the effect estimate was 4.19, and it was somewhat reduced with narrower CIs. The effect size also decreased to 0.21. When using the half-size bandwidth, it resulted in a minor variation in the ITT estimate, with effect size differences of about 0.02. None of sensivity analysis results reached statistical significance with respect to the conventional or robust confidence intervals. Overall, there is insufficient evidence to support a positive effect of students being assigned to the state's pre-K program at the cutoff.

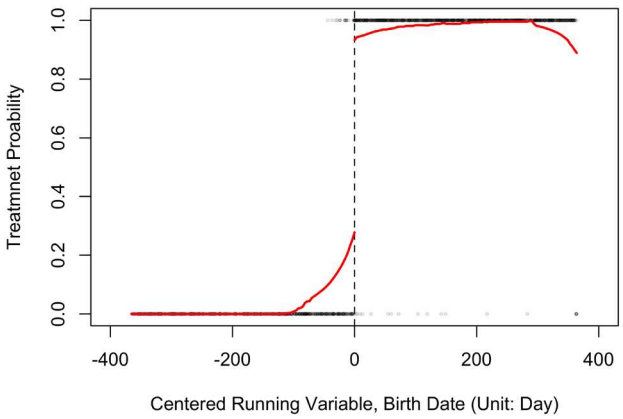


Fig. 5 Discontinuity in the treatment probability at the cutoff

LATE at the cutoff

Table 3 presents the compliance rates for New Jersey's pre-K program based on the eligibility and receipt statuses of children. In our study sample, approximately 56.3% of the students (1,122 students) were eligible for the pre-K program. Among those who were eligible, around 98.1% actually received the program, while 97.8% of the ineligible students did not receive it. This yields a non-compliance rate of about 2%, which is very small. When focusing on observations near cutoff, such as within an optimal bandwidth of 69.4 days, the compliance rate is further reduced to about 1%.

Furthermore, Fig. 5 displays a visual plot of the RD design for the treatment probability at the cutoff using local linear regression. The figure reveals an evident discontinuity at the cutoff, although it is less than 1 due to noncompliance cases.

Table 4 summarizes the LATE results for different bandwidth choices, with and without additional covariate adjustment. Similar to ITT estimates, effect size estimates are also provided. The LATE estimates were larger than the corresponding ITT estimates. Note that the LATE is always larger than the ITT effect on an absolute scale because the LATE estimate is calculated as the ITT estimate divided by the estimated difference in the probability of receiving treatment between either side of the cutoff. Specifically, using an optimal bandwidth of 69.4 days, the LATE estimate

Table 4 Local average treatment effect (LATE) of New Jersey's pre-K program at the cutoff

	Bandwidth	Estimate	Effect Size	95% CI	95% Robust CI
LATE (Optimal BW)	69.4	9.57	0.48	[−1.31, 20.45]	[−2.09, 22.42]
Optimal BW w/ Covs	69.4	4.34	0.22	[−6.50, 15.17]	[−14.35, 17.88]
Half-BW	34.7	8.64	0.44	[−7.04, 24.33]	[−16.77, 24.05]

Effect sizes are calculated using the empirical sample standard deviation of the outcome

CI represents the confidence interval, *BW* represents the bandwidth, and *Covs* represents covariate adjustment

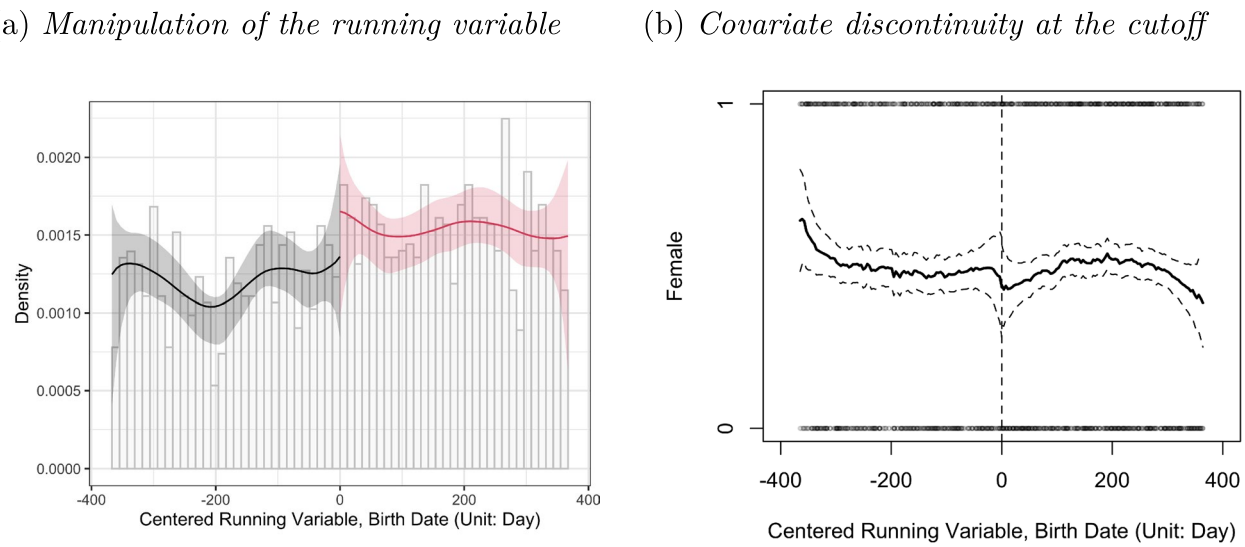


Fig. 6 Results of falsification tests

was 9.57 points, but it was not statistically significant. The effect size associated with this estimate was 0.48, indicating a small to medium effect. Incorporating additional covariate adjustments improved efficiency of the LATE estimate, but reduced the LATE estimate to 4.34 with an effect size of 0.22, which is about half of the original effect size estimate. The differences in LATE estimates without and with covariates may be due to potential attribution or selection bias between either side of the cutoff within the window. This observation is further discussed in the next subsection below. When using a different bandwidth, reducing it by half only resulted in a decrease of less-than one point in the LATE estimate, with effect size differences of about 0.04. Overall, there is insufficient evidence to support a positive effect of receiving the state's pre-K program among compliers at the cutoff, although the point estimates and effect sizes are larger compared to those of the ITT effect.

Falsification tests

We conducted a set of falsification tests to (i) detect whether the running variable is manipulated at the cutoff (i.e., whether the density of the running variable is discontinuous at the cutoff) and (ii) evaluate whether units from different sides of the cutoff have differences in measured covariates at the cutoff (i.e., whether the measured covariates are discontinuous at the cutoff). Figure 6a visualizes the results of the manipulation test using the local polynomial density estimator proposed by Cattaneo et al. (2020), an improved version of McCrary's (2008) test. The figure suggests that the level to the left of the cutoff (i.e., the control condition) is lower than that to the right (i.e., the treatment condition), which could indicate potential differential attrition issues between the two sides of the cutoff. However, the confidence bands overlap at the cutoff, and this indicates that the observed discontinuity at this point is not statistically significant. Furthermore, the numerical results from the manipulation test using the estimator proposed by Cattaneo et al. (2020) show insignificance (test statistic = 0.3177, p-value = 0.7507), in contrast to the significant result obtained from the McCrary (2008) test (test statistic = 2.1137, p-value = 0.0345). While we may rely on the former test due to its improved development, we should still exercise caution when interpreting RD results, given the visual inspection results that potentially

indicate differential attrition or sample selection around the cutoff.

To examine the discontinuity of measured covariates at the cutoff, we analyzed the relationship between the running variable and each covariate in the data. Figure 6b presents the result for the gender variable (female = 1, male = 0). The plot includes a local linear regression line with confidence intervals, and there is no clear evidence of a discontinuity in the gender proportion at the cutoff. We performed similar analyses for other covariates and found no indications of discontinuity at the cutoff. As a result, we are not concerned about non-random sorting near the cutoff from measured covariates. However, it is important to note that there remains a possibility of non-random sorting with respect to unmeasured covariates (e.g., pre-test).

Conclusions

In this paper, we have provided a comprehensive review of traditional RD designs and recent developments, particularly those that are more related to educational contexts. Our empirical analysis in Sect. Educational example: New Jersey's pre-K programs demonstrates specific instructions on the RD designs. To conclude, we offer practical guidelines for utilizing RD designs:

1. Plot the data using scatterplots and summary smoothing lines to visualize the relationship between the running variable and the outcome (see Fig. 4). Adjust the regression lines with appropriate bandwidths and degrees of polynomial to assess local discontinuity at the cutoff.
2. Estimate the treatment effect using nonparametric or parametric regression. For nonparametric regression, determine an optimal bandwidth that minimizes the MSE of the RD effect estimator and use robust standard errors or confidence intervals (see Table 2). For parametric regression, assess the appropriate degrees of polynomials based on F-tests or the AIC.
3. Conduct the sensitivity of the RD treatment effects by varying the bandwidths (for nonparametric regression) and including measured covariates in the regression models (see Table 2).
4. Assess threats to validity regarding manipulation or sorting near the cutoff (see Sect. Falsification tests). Conduct visual inspections and perform formal statistical tests (e.g., the manipulation test using Cattaneo et al. (2020)).

   For the fuzzy RD design with imperfect compliance, we adhere to the above guidelines for examining the ITT at the cutoff. To assess the LATE at the cutoff, we provide additional guidelines:

5. Summarize the compliance rates (see Table 3) and plot the relationship between the running variable and treatment (see Fig. 5), in addition to the RD plot with outcome.
6. Estimate the treatment effect using parametric TSLS regression or nonparametric approaches (see Table 4). Use the same bandwidth and order of polynomials for both the treatment regression and outcome regression.

Overall, although this review paper does not cover every aspect of RD designs due to space limitations, we hope that our review and practical guidelines will be valuable for researchers who are interested in applying RD designs to evaluate education policies or programs.

Acknowledgements We would like to thank two editors for this special issue, Peter Steiner and Yongnam Kim, as well as an anonymous reviewer, for their useful comments that have improved the manuscript. We would also like to thank Vivian Wong for granting permission to use her data from {wong et al. 2007} for the purpose of demonstrating regression discontinuity designs in this work.

Funding This work was partly supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those AERA or NSF.

Declarations

Conflicting interests The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association, 110*(512), 1331–1344. https://doi.org/10.1080/01621459.2015.1012259

Bergolo, M., & Galván, E. (2018). Intra-household behavioral responses to cash transfer programs. evidence from a regression discontinuity design. *World Development, 103*, 100–118. https://doi.org/10.1016/j.worlddev.2017.10.030

Brunner, E. J., Dougherty, S. M., & Ross, S. L. (2023). The effects of career and technical education: Evidence from the connecticut technical high school system. *Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_01098

Bulus, M. (2021). Minimum detectable effect size computations for cluster-level regression discontinuity studies: Specifications beyond the linear functional form. *Journal of Research on Educational Effectiveness, 15*(1), 151–177. https://doi.org/10.1080/19345747.2021.1947425

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2019). Regression discontinuity designs using covariates. *The Review of Economics and Statistics, 101*(3), 442–451. https://doi.org/10.1162/rest_a_00760

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2023). Rdrobust: Robust data-driven statistical inference in

regression-discontinuity designs [R package version 2.2]. https://CRAN.R-project.org/package=rdrobust.
Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs: Robust nonparametric confidence intervals. *Econometrica, 82*(6), 2295–2326. https://doi.org/10.3982/ecta11757
Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65*(2), 684. https://doi.org/10.2307/1131410
Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica, 83*(6), 2453–2483. https://doi.org/10.3982/ecta11224
Card, D., Lee, D. S., Pei, Z., & Weber, A. (2017). Regression kink design: Theory and practice. *Advances in econometrics* (pp. 341–382). Emerald Publishing Limited. https://doi.org/10.1108/s0731-905320170000038016
Carlson, D., & Knowles, J. E. (2016). The effect of english language learner reclassification on student ACT scores, high school graduation, and postsecondary enrollment: Regression discontinuity evidence from wisconsin. *Journal of Policy Analysis and Management, 35*(3), 559–586. https://doi.org/10.1002/pam.21908
Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press. https://doi.org/10.4135/9781412993869
Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press. https://doi.org/10.4135/9781412993869
Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference, 3*(1), 1–24. https://doi.org/10.1515/jci-2013-0010
Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal: Promoting communications on statistics and Stata, 18*(1), 234–261. https://doi.org/10.1177/1536867x1801800115
Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association, 115*(531), 1449–1455. https://doi.org/10.1080/01621459.2019.1635480
Cattaneo, M. D., Jansson, M., & Ma, X. (2023). Rddensity: Manipulation testing based on density discontinuity [R package version 2.4]. https://CRAN.R-project.org/package=rddensity.
Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2020). Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *Journal of the American Statistical Association, 116*(536), 1941–1952. https://doi.org/10.1080/01621459.2020.1751646
Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2019). Power calculations for regression-discontinuity designs. *The Stata Journal: Promoting communications on statistics and Stata, 19*(1), 210–245. https://doi.org/10.1177/1536867x19830919
Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics, 14*(1), 821–851. https://doi.org/10.1146/annurev-economics-051520-021409
Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality (B. S. Barnow, Ed.). *Journal of Policy Analysis and Management, 36*(3), 643–681. https://doi.org/10.1002/pam.21985
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. https://doi.org/10.1016/j.jeconom.2007.05.002
Coyne, M. D., Oldham, A., Dougherty, S. M., Leonard, K., Koriakin, T., Gage, N. A., Burns, D., & Gillis, M. (2018). Evaluating the effects of supplemental reading intervention within an MTSS or RTI reading reform initiative using a regression discontinuity design. *Exceptional Children, 84*(4), 350–367. https://doi.org/10.1177/0014402918772791
Dong, Y. (2014). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics, 30*(3), 422–446. https://doi.org/10.1002/jae.2369
Dong, Y., & Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics, 97*(5), 1081–1092. https://doi.org/10.1162/rest_a_00510
Figlio, D., Holden, K. L., & Ozek, U. (2018). Do students benefit from longer school days? regression discontinuity evidence from florida's additional hour of literacy instruction. *Economics of Education Review, 67*, 171–183. https://doi.org/10.1016/j.econedurev.2018.06.003
Figlio, D., & Özek, U. (2023). The unintended consequences of test-based remediation (tech. rep.). NBER Working Paper No. w30831. https://doi.org/10.2139/ssrn.4320579.
Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations*. University of Wisconsin-Madison.
Grembi, V., Nannicini, T., & Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics, 8*(3), 1–30. https://doi.org/10.1257/app.20150076
Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201–209. https://doi.org/10.1111/1468-0262.00183
Hausman, C., & Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics, 10*(1), 533–552. https://doi.org/10.1146/annurev-resource-121517-033306
Heissel, J. A., & Ladd, H. F. (2018). School turnaround in north carolina: A regression discontinuity analysis. *Economics of Education Review, 62*, 302–320. https://doi.org/10.1016/j.econedurev.2017.08.001
Hjalmarsson, R. (2009). Juvenile jails: A path to the straight and narrow or to hardened criminality? *The Journal of Law and Economics, 52*(4), 779–809. https://doi.org/10.1086/596039
Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies, 79*(3), 933–959. https://doi.org/10.1093/restud/rdr043
Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001
Imbens, G., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/cbo9781139025751
Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis, 23*(1), 127–155. https://doi.org/10.1093/pan/mpu014
Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics, 142*(2), 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004
Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142*(2), 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003
Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281
Lee, M. G., & Soland, J. G. (2022). Does reclassification change how english learners feel about school and themselves? evidence from a regression discontinuity design. *Educational Evaluation and Policy Analysis, 45*(1), 27–51. https://doi.org/10.3102/01623737221097419

Li, F., Mercatanti, A., Mäkinen, T., & Silvestrini, A. (2021). A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds. *The Annals of Applied Statistics, 15*(1), 304–322. https://doi.org/10.1214/20-AOAS1396

Linden, A., & Adams, J. L. (2012). Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice, 18*(2), 317–325. https://doi.org/10.1111/j.1365-2753.2011.01768.x

Ludwig, J., & Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics, 122*(1), 159–208. https://doi.org/10.1162/qjec.122.1.159

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

McEachin, A., Domina, T., & Penner, A. (2020). Heterogeneous effects of early algebra across california middle schools. *Journal of Policy Analysis and Management, 39*(3), 772–800. https://doi.org/10.1002/pam.22202

Mealli, F., & Rampichini, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *Journal of the Royal Statistical Society Series A: Statistics in Society, 175*(3), 775–798. https://doi.org/10.1111/j.1467-985x.2011.01022.x

Melguizo, T., Sanchez, F., & Velasco, T. (2015). Credit for low-income students and access to and academic performance in higher education in colombia: A regression discontinuity approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2608642

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9 (with discussion). *Statistical Science, 4*, 465–480.

Nielsen, H. S., Sørensen, T., & Taber, C. (2010). Estimating the effect of student aid on college enrollment: Evidence from a government grant policy reform. *American Economic Journal: Economic Policy, 2*(2), 185–215. https://doi.org/10.1257/pol.2.2.185

Nomi, T., & Raudenbush, S. W. (2016). Making a success of algebra for all. *Educational Evaluation and Policy Analysis, 38*(2), 431–451. https://doi.org/10.3102/0162373716643756

Otsu, T., Xu, K.-L., & Matsushita, Y. (2013). Estimation and inference of discontinuity in density. *Journal of Business & Economic Statistics, 31*(4), 507–524. https://doi.org/10.1080/07350015.2013.818007

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness, 5*(1), 83–104. https://doi.org/10.1080/19345747.2011.609583

Rokkanen, M. (2015). Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design (tech. rep.). Discuss. Pap. 1415-03, Dep. Econ., Columbia Univ., New York.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962. https://doi.org/10.2307/2289065

Schochet, P. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics, 34*(2), 238–266.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. Standards for regression discontinuity designs. 2010. http://ies.ed.gov/ncee/wwc/pdf/wwc%5C_rd.pdf.

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. *Journal of Public Economics, 152*, 154–169. https://doi.org/10.1016/j.jpubeco.2017.06.004

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

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 236–258). Oxford University Press.

Steiner, P. M., Kim, Y., Hall, C. E., & Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological Methods & Research, 46*(2), 155–188. https://doi.org/10.1177/0049124115582272

Suk, Y., & Kim, Y. (2023). Fuzzy regression discontinuity designs with multiple control groups under one-sided noncompliance: Evaluating extended time accommodations. https://doi.org/10.31234/osf.io/sa96g

Suk, Y., Steiner, P. M., Kim, J.-S., & Kang, H. (2022). Regression discontinuity designs with an ordinal running variable: Evaluating the effects of extended time accommodations for english-language learners. *Journal of Educational and Behavioral Statistics, 47*(4), 459–484. https://doi.org/10.3102/10769986221090275

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 Regression discontinuity designs (pp. 237–279). Emerald Publishing Limited. https://doi.org/10.1108/s0731-905320170000038011.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*(6), 309–317. https://doi.org/10.1037/h0044319

Villamizar-Villegas, M., Pinzon-Puerto, F. A., & Ruiz-Sanchez, M. A. (2021). A comprehensive history of regression discontinuity designs: An empirical survey of the last 60 years. *Journal of Economic Surveys, 36*(4), 1130–1178. https://doi.org/10.1111/joes.12461

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. https://doi.org/10.1002/pam.21721

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2007). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27*(1), 122–154. https://doi.org/10.1002/pam.20310

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables. *Journal of Educational and Behavioral Statistics, 38*(2), 107–141. https://doi.org/10.3102/1076998611432172