Methods for Policy Analysis

Burt S. Barnow, Editor

COMPARING INFERENCE APPROACHES FOR RD DESIGNS: A REEXAMINATION OF THE EFFECT OF HEAD START ON CHILD MORTALITY

Matias D. Cattaneo, Rocío Titiunik, and Gonzalo Vazquez-Bare

Abstract

The regression discontinuity (RD) design is a popular quasi-experimental design for causal inference and policy evaluation. The most common inference approaches in RD designs employ "flexible" parametric and nonparametric local polynomial methods, which rely on extrapolation and large-sample approximations of conditional expectations using observations somewhat near the cutoff that determines treatment assignment. An alternative inference approach employs the idea of local randomization, where the very few units closest to the cutoff are regarded as randomly assigned to treatment and finite-sample exact inference methods are used. In this paper, we contrast these approaches empirically by re-analyzing the influential findings of Ludwig and Miller (2007), who studied the effect of Head Start assistance on child mortality employing parametric RD methods. We first review methods based on approximations of conditional expectations, which are relatively well developed in the literature, and then present new methods based on randomization inference. In particular, we extend the local randomization framework to allow for parametric adjustments of the potential outcomes; our extended framework substantially relaxes strong assumptions in prior literature and better resembles other RD inference methods. We compare all these methods formally, focusing on both estimands and inference properties. In addition, we develop new approaches for randomization-based sensitivity analysis specifically tailored to RD designs. Applying all these methods to the Head Start data, we find that the original RD treatment effect reported in the literature is quite stable and robust, an empirical finding that enhances the credibility of the original result. All the empirical methods we discuss are readily available in general purpose software in R and Stata; we also provide the dataset and software code needed to replicate all our results.

Journal of Policy Analysis and Management, Vol. 36, No. 3, 643–681 (2017)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
Supporting Information is available in the online issue at wileyonlinelibrary.com.
DOI:10.1002/pam.21985

INTRODUCTION

Every year, federal governments throughout the world spend large fractions of their budgets on programs aimed at assisting low-income populations in obtaining health care, housing, food, and education. In the United States, the total federal spending on the ten largest means-tested programs and tax credits for low-income households rose from 1 percent to 4 percent of the GDP between 1972 and 2012, recently totaling nearly 600 billion dollars (CBO, 2013). Given the amount of resources devoted to such programs, and the importance of their goals, evaluating whether they achieve their intended policy objectives is crucial. The most compelling method of program evaluation is one that randomly assigns units to receive (or not receive) the program benefits. However, in the context of means-tested social programs, experimental designs are rarely feasible unless the number of applicants exceeds the number of available program slots; otherwise, the creation of an experimental control group would have the unethical consequence of leaving some program slots unfilled. Thus, the use of non-experimental research designs to evaluate social programs is often unavoidable.

This article focuses on the regression discontinuity (RD) design, a non-experimental strategy that allows researchers to obtain a valid "quasi-experimental" control group when the treatment of interest is not randomized—including situations when the neediest applicants receive the program first. This design is based on two main assumptions. The first is that each program applicant receives a score, and the program is given to all applicants whose score exceeds a known cutoff (the treatment group), and withheld from all applicants whose score is lower than the cutoff (the control group). This feature is common to all RD designs with perfect compliance and is easily verifiable, since it refers to an observable treatment assignment mechanism that is usually set ex ante by the institution granting the program. The second assumption is that units in the control and treatment groups near the cutoff are valid counterfactuals of each other, ruling out program participants' ability to precisely manipulate their score value and hence their treatment status. This assumption is sometimes described as ruling out the "endogenous sorting" of units around the cutoff. For early reviews and general methodological discussions see, for example, Cook (2008), Imbens and Lemieux (2008), Lee and Lemieux (2010), and Wing and Cook (2013). See also Cattaneo and Escanciano (2017) for an edited volume with very recent overviews, discussions, and references.

We discuss two approaches to the analysis of RD designs, each of which adopts a different version of the non-sorting assumption. In the first framework, valid counterfactuals follow from the assumption that the conditional expectations of potential outcomes given the score are continuous at the cutoff, ensuring that the characteristics of treated participants with scores very near the cutoff are not abruptly different from the characteristics of control participants whose scores are also close to the cutoff. This continuity assumption leads to nonparametric identification of a local average treatment effect (ATE) at the cutoff, and justifies the use of (nonparametric) local polynomial techniques for estimation and inference (Hahn, Todd, & van der Klaauw, 2001). In the second framework, the validity of the treatment-control comparisons follows from assuming that the treatment is as-if randomly assigned in a small window around the cutoff. This local randomization assumption, which is stronger than the continuity assumption invoked by the first framework, justifies the use of methods from the analysis of experiments literature for estimation and inference (Cattaneo, Frandsen, & Titiunik, 2015). This second approach is motivated by the influential work of Lee (2008), who discussed the idea of RD designs as local randomized experiments (see also the original paper of Thistlethwaite & Campbell, 1960).

In many empirical applications of RD designs, researchers often combine both frameworks in an informal way—for example, using "flexible" parametric polynomial methods to estimate treatment effects but using local randomization methods to provide empirical evidence in favor of the design and heuristic causal interpretations. Our main goal in this article is to formalize and discuss the differences between these methods, both methodologically and empirically, employing as a case study the influential findings of Ludwig and Miller (2007), who studied the effect of Head Start assistance on child mortality employing parametric RD methods. For very recent related work see also Pihl (2016). We compare and contrast the two RD methodological approaches, in each case discussing formally how to define and interpret the parameter of interest and how to perform estimation and inference.

In the continuity-based framework, we discuss the two most common estimation and inference methods: (i) global polynomial and flexible parametric inference, and (ii) nonparametric local polynomial inference. We advise against the first set of methods, since they are parametric in nature, and de facto ignore the impact of over-fitting higher-order polynomials for boundary estimation (global approach) or parametric misspecification bias and neighborhood selection (flexible parametric approach). We recommend the second set of methods because—instead of imposing a parametric model—they rely on nonparametric approximations of the unknown regression functions on either side of the cutoff, and employ nonparametric large-sample inference techniques for estimation and inference that take into account the error in the approximation (i.e., misspecification error or smoothing bias). In this framework, the population parameter of interest might not be regarded as causal but is, nonetheless, policy relevant and very useful (though local in nature without additional assumptions).

In the local randomization framework, where treatment assignment is assumed to be as-if randomly assigned in a small window around the cutoff, we propose to employ exact randomization-based inference methods, first formally developed in the RD setting by Cattaneo et al. (2015). The reason is that the window where this local randomization assumption is plausible is likely to be small and thus contain few observations, which may render large-sample approximation methods invalid. Randomization-based methods avoid this problem because they lead to inferences that are finite-sample correct.[1] We make two novel methodological contributions to the local randomization framework. First, we extend the local randomization RD framework to develop a formal model of transformed outcomes based on flexible parametric adjustments of the potential outcomes; this allows the potential outcomes to depend on the running variable in a flexible way and therefore substantially relaxes strong assumptions previously employed in the literature (e.g., that the average response of the outcome near the cutoff is constant). In particular, our novel methodology gives a formal justification for parametric polynomial fitting near the cutoff before employing randomization inference techniques. Second, we propose new randomization-based sensitivity methods specifically developed for RD designs. Both contributions are new to the literature, building, and improving on the randomization-inference framework proposed in Cattaneo et al. (2015).

We illustrate our discussion with an analysis of Head Start, a U.S. federal program that provides education, health, nutritional, and social services for children from birth to age five, including center-based preschool services for three-year-olds and four-year-olds (Head Start Report, 2010). In particular, we re-examine the influential article by Ludwig and Miller (2007), who employed an RD design to study the

[1] For textbook reviews of randomization-based methods see Rosenbaum (2002b, 2010) and Imbens and Rubin (2015). Recent applications of randomization inference methods to the social sciences include Imbens and Rosenbaum (2005), Ho and Imai (2006), Bowers, Fredrickson, and Panagopoulos (2013), and Keele, Titiunik, and Zubizarreta (2015).

effects of Head Start on child mortality, relying on a discontinuity on access to program funds that occurred in 1965 when the program was first implemented. Specifically, in order to ensure that applications from the poorest communities would be represented in a nationwide grant competition for the program's funds, the federal government provided assistance to the 300 poorest counties in the United States to write and submit applications for Head Start funding. This led to increased Head Start participation and funding rates in these counties, creating a discontinuity in program participation at the 300th poorest county that can be used to estimate an RD treatment effect of the program. Using flexible parametric methods, Ludwig and Miller (2007) found that access to increased Head Start funding decreased 1973 to 1983 county-level mortality rates of children of age five to nine due to causes affected by Head Start's health services component. The mortality reduction reported is large, from approximately 3.2 to 1.9 deaths per 100,000.

Our re-examination of the Head Start data shows that this reduction in child mortality is robust to different RD assumptions and estimation strategies. Adopting a continuity-based framework, the global and flexible parametric approaches yield a point estimate of about −2.5 deaths per 100,000 (statistically significant at 5 percent level), and the robust local nonparametric approach leads to a very similar conclusion, with an RD treatment effect of about −2.3 deaths per 100,000 (statistically significant at 5 percent level). Adopting a local randomization framework leads to similar conclusions: we estimate treatment effects of −2.3 and −2.5 deaths per 100,000, we reject the (sharp) null hypothesis that the treatment has no effect for any unit with randomization-based p-values below 0.01, and we show that these findings are robust to window selection, parametric misspecification, local interference, and the presence of unobserved confounders assessed with our newly developed sensitivity analysis methods. All the methods are implemented in publicly available R and Stata software packages (see Calonico et al., 2017; Calonico, Cattaneo, & Titiunik, 2014a, 2015b; Cattaneo, Titiunik, & Vazquez-Bare, 2016). Accompanying this article, we also provide data and complete replication codes in both R and Stata. Latest software, data, and codes are publicly available at https://sites.google.com/site/rdpackages/.

The rest of this article is organized as follows. The next section introduces the basic RD setup and presents the Head Start data. The following sections offer a discussion of the continuity-based framework, employing global and local parametric techniques to estimate Head Start effects; introduce the RD local randomization framework and present our new methodological developments, including flexible outcome adjustments and novel sensitivity methods, and apply them to the Head Start data; collect all our empirical findings and provide a methodological (and substantive) discussion; and offer comprehensive recommendations for practice and a conclusion. The Supporting Information Appendix reports additional methodological discussions and empirical results, and presents an extension of our new randomization-based methods to the case of "fuzzy" RD designs where compliance is imperfect.[2] See Ganong and Jäger (2016) for permutation-based inference in "kink" RD designs (Card et al., 2015; Chiang & Sasaki, 2016).

BASIC RD SETUP: HEAD START PROGRAM

We start by presenting the main features of the RD design that are common to both the continuity-based and the local randomization framework. We focus on the so-called sharp RD design, where treatment compliance is perfect or the researcher

[2] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

focuses on the intention-to-treat parameter—as occurs in the Head Start empirical application.

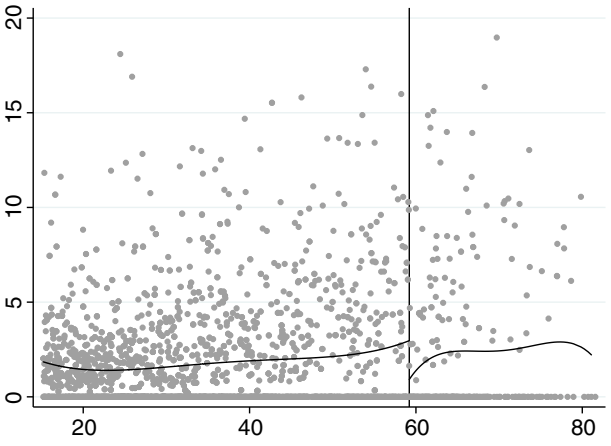
In the sharp RD design, treatment assignment is a deterministic function of the running variable or score: each unit with observed running variable R_i below the known threshold $\bar{r}$ is assigned to the control group ($D_i = 0$) and each unit with $R_i \geq \bar{r}$ is assigned to the treatment group ($D_i = 1$). Thus, $D_i = 1(R_i \geq \bar{r})$ for each unit i in the sample, with $1(\cdot)$ denoting the indicator function. The running variable R_i is assumed to be random throughout, as it determines treatment assignment for each unit in the sample. In the case study of Head Start (HS), the sample consists of U.S. counties, $i = 1, 2, \ldots, n$ with $n = 2,804$, and the score is a county-level poverty index constructed in 1965 by the federal government based on 1960 census information, with support $R_i \in [15.2085, 93.0717]$. The cutoff is $\bar{r} = 59.1984$ and $D_i = 1(R_i \geq 59.1984)$, which was chosen so that the number of treated counties would be exactly 300.

We employ the potential outcomes framework to analyze all approaches to RD analysis in a unified way. In this framework, each unit is assumed to have one potential or underlying outcome for all or some subset of all possible treatment assignments, and the treatment effects are defined in terms of these underlying outcomes, which are distinguished from the observed outcomes. Depending on the approach to inference considered, these potential outcomes may be either random or fixed quantities. Below, we will use $y_i(\cdot)$ to denote fixed potential outcomes and $Y_i(\cdot)$ to denote random potential outcomes for each unit i in the sample (the actual evaluation variables depend on the framework, as we discuss below). When the potential outcomes are modeled as random, they are seen as a sample from some underlying (super) population; this model is particularly useful for some frequentist large-sample arguments. When the potential outcomes are modeled as fixed quantities of the units in the sample, then the operating assumption is that the sample is the population of interest, and inference is based on the randomization mechanism and is valid only for those observed units. The latter assumption is most common in the analysis of experiments and related randomization inference frameworks (Imbens & Rubin, 2015; Rosenbaum, 2002b, 2010).

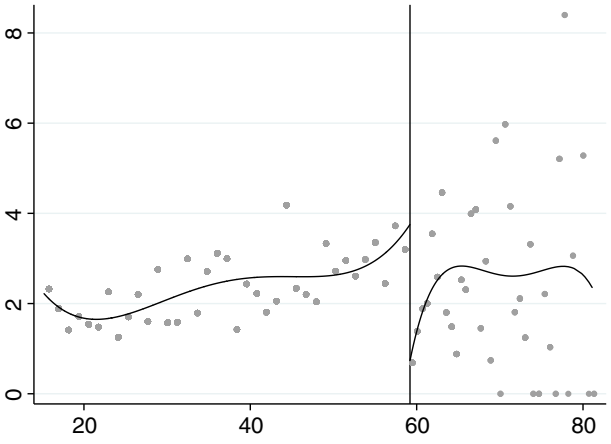
Graphical Presentation

One of the main advantages of the RD design is related to the ease with which it can be visualized and assessed intuitively. We begin by presenting the RD design graphically using the Head Start data. Figure 1 plots the county-level death rates of children ages five to nine as a function of the county's poverty index, using the data first used by Ludwig and Miller (2007) to study the impact of Head Start on child mortality during the 1973 to 1983 period. Figure 1a plots the raw mortality counts, while Figure 1b plots binned means of child mortality with evenly-spaced bins chosen optimally to mimic the variability of the outcome variable (Calonico, Cattaneo, & Titiunik, 2015a). The solid lines are fourth-order global polynomial fits, and the vertical line indicates the value of the poverty index cutoff that determined technical assistance to apply for Head Start funding. The smoothed plot in Figure 1b illustrates a downward jump right at the poverty index cutoff, with mortality being lower immediately to the right of the cutoff in counties that qualified to receive assistance and higher immediately to the left in counties where no assistance was offered and Head Start participation was much lower.

Figure 1b already hints heuristically to a potential RD treatment effect of Head Start on child mortality for counties having a 1960 poverty index of about $\bar{r} = 59.1984$. In the rest of the paper, we formalize this heuristic finding, discussing the different features of population parameters and inference methods considered.



(a) Scatter Plot, Raw Data, $N^- = 2,455$, $N^+ = 290$



(b) RD Plot, ES, and MV, $J_- = 37$, $J_+ = 38$

Notes: (i) In panel (a), N^- and N^+ denote the sample sizes for control and treatment units, respectively; (ii) in panel (b) bins are evenly spaced (ES) and their total number (J_-, J_+) chosen to mimic variance (MV); (iii) solid blue lines depict fourth-order polynomial fits using control and treated units separately; and (iv) dots depict raw data points in panel (a) and sample average of outcome variable within each bin in panel (b).

Figure 1. Scatter and RD Plot. Head Start Data.

Falsification Tests

The key idea underlying the RD design is that because units (counties in the Head Start case) do not have precise control over their running variable, their characteristics (both observable and unobservable) should not change abruptly at the cutoff, leading to comparable control and treatment groups. In particular, lack of systematic sorting around the cutoff will be compatible with a continuous density of the running variable near the cutoff and continuous conditional expectation functions of potential outcomes near the cutoff. A key issue in any empirical study employing RD methods is to assess the validity of the design by providing evidence

in favor of these assumptions. We consider the three most common approaches for the falsification of RD designs.

1. **Continuity away from cutoff**. This approach seeks to check graphically—and later formally if needed—whether the outcome variables exhibit discontinuities over the support of the running variable at places other than the actual cutoff $\bar{r}$. The approach is based on repeated sampling assumptions and studies the behavior of the conditional expectations $\mathbb{E}[Y_i(0)|R_i]$ and $\mathbb{E}[Y_i(1)|R_i]$ over the range of the running variable R_i. See Calonico et al. (2015a) for further details and discussion. In addition, see Cerulli et al. (2017) for a related approach, building on Dong and Lewbel (2015), which looks at the local (to the cutoff) sensitivity of the treatment effect estimate.
2. **Running variable manipulation**. This approach was originally proposed by McCrary (2008) and builds on the idea that units in the sample should not be able to sort precisely around the cutoff (i.e., no self-selection into control or treatment status). Thus, in the absence of precise manipulation of their running variable, placement of units just below (control group) and just above (treatment group) should be as-if random near the RD cutoff $\bar{r}$, and the number of treated and control units in a neighborhood of the cutoff should be approximately similar. This idea leads to falsification methods based on comparing the number or "density" of treated and control units near the cutoff. See Frandsen (2017) for a related approach when the running variable takes discrete values (which is not the case in the Head Start application), and Jales and Yu (2017) for a review of related approaches exploiting a discontinuity in density.
3. **Placebo treatment effects**. This approach was formalized by Lee (2008) and builds on the idea that pre-intervention covariates and post-treatment outcomes on which the treatment is known to have no effect (also called "placebo" outcomes) should exhibit a zero RD treatment effect. This method is implemented by conducting inference on RD treatment effects using both types of variables as outcomes at the true cutoff value $\bar{r}$ and, sometimes, at artificial cutoff points on the support of the running variable. The latter gives a formal test for detecting potential discontinuities over the support of the running variable, as discussed above.

We perform falsification tests to check the validity of the RD design employing all three approaches. We present the main results here, but relegate details to the Supporting Information Appendix to conserve space.[3] To check for continuity away from the cutoff on the outcome variable, we construct RD plots using two different approximations to the regression functions, following the recommendations in Calonico et al. (2015a). These figures are shown in the Supporting Information Appendix. Employing the Head Start data, we found no empirical evidence of discontinuities away from the real RD cutoff $\bar{r} = 59.1984$.

The first falsification approach described above is graphical and global in nature, as it looks at the behavior of the data over the full support of the running variable. In contrast, the second falsification approach focuses on the observations near the cutoff and employs only the distribution of the running variable—that is, it does not employ outcomes or covariates. The idea of manipulation of the running variable is quite important in RD designs; in the Head Start application, it is highly unlikely ex ante because assistance was offered to counties in 1965 based on their poverty index R_i constructed using the 1960 census data—it is unlikely that in 1960 counties could have anticipated that a policy would be based on whether the poverty index

[3] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Table 1. Falsification test based on the running variable R_i.

	Binomial tests		
h	N_W^-	N_W^+	p-value
0.3	9	10	1.000
0.5	18	16	0.864
0.7	24	22	0.883
0.9	32	27	0.603
1.1	43	33	0.302
1.3	51	38	0.203

Notes: (i) Cutoff is $\bar{r} = 59.1984$ and $W = [\bar{r} - h, \bar{r} + h]$denotes the window around the cutoff used for each choice of bandwidth; (ii) binomial test p-values are computed using exact binomial distribution with probability $q = 1/2$.

was exactly $\bar{r} = 59.1984$ five years later. Nonetheless, it is interesting to illustrate how to assess the plausibility of the RD design using this second method.

We implement the running variable manipulation falsification approach in two distinct ways. First, we employ a binomial test aimed to formally check whether the number of observations in the control and treatment groups near the cutoff is surprisingly different from the number that would be expected in a random sample of Bernoulli trials with a pre-specified probability $q \in (0, 1)$. This first implementation is based on the idea that if units within a window or neighborhood $W = [\bar{r} - h, \ \bar{r} + h]$ around the cutoff were randomly assigned to treatment with probability q, then the number of effective control units $N_W^- = \sum_{i=1}^n 1(\bar{r} - h \leq R_i < \bar{r})$ and effective treatment units $N_W^+ = \sum_{i=1}^n 1(\bar{r} \leq R_i \leq \bar{r} + h)$ should follow a binomial distribution, a fact that can be tested empirically. Here h controls the width of the neighborhood W around the RD cutoff—this tuning parameter plays a crucial role in the RD literature, as we discuss throughout this paper. To implement this idea, we first choose a neighborhood or window W near the cutoff $\bar{r}$ where this test is carried out, and a probability of treatment assignment q. In practice, in the absence of additional information, $q = 1/2$ is the most natural choice. Table 1 shows the results of the binomial test for a few small windows near the cutoff—in the Supporting Information Appendix we present a more complete analysis.[4] The empirical findings are consistent with what would be observed under a simple Bernoulli assignment mechanism in small windows near the cutoff.

An alternative implementation of this falsification method is applicable to RD designs where the running variable is a continuous random variable—as it occurs in our Head Start application—and is based on the idea of continuity of the (Lebesgue) density of R_i near the cutoff. This idea was originally proposed by McCrary (2008), and is implemented by conducting a formal nonparametric hypothesis test of continuity of the probability density function of R_i at $\bar{r}$. We implement this falsification test employing the recent results in Cattaneo, Jansson, and Ma (2016a, 2016b), which rely on local polynomial distribution regression methods, bias-correction techniques, and robust distributional approximations to conduct the hypothesis test. For brevity, we refer the reader to the latter references for further details. The results are presented in Table 2, where we fail to reject the null hypothesis that the density of the running variable is continuous at the cutoff and thus obtain additional empirical evidence in favor of the validity of the RD design in the Head Start empirical application.

[4] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Table 2. Falsification test based on the running variable R_i.

	Density tests				
	h_-	h_+	N_W^-	N_W^+	p-value
Method					
Unrestricted, 2-h	10.151	9.213	351	221	0.788
Unrestricted, 1-h	9.213	9.213	316	221	0.607
Restricted (1-h)	13.544	13.544	482	255	0.655

Notes: (i) Cutoff is $\bar{r} = 59.1984$ and $W = [\bar{r} - h, \bar{r} + h]$ denotes the symmetric window around the cutoff used for each choice of bandwidth; (ii) Density test p-values are computed using Gaussian distributional approximation to bias-corrected local-linear polynomial estimator with triangular kernel and robust standard errors; (iii) column "Method" reports unrestricted inference with two distinct estimated bandwidths ("U, 2-h"), unrestricted inference with one common estimated bandwidth ("U, 1-h"), and restricted inference with one common estimated bandwidth ("R, 1-h"). See Cattaneo, Jansson, and Ma (2016a, 2016b) for methodological and implementation details.

We postpone the discussion of the third falsification method based on placebo treatment effects to the following sections, as this method requires estimation and inference techniques for RD analysis. We turn to discussion of these techniques below.

RD BASED ON CONTINUITY AT THE CUTOFF

The first approach to RD analysis that we consider is one that assumes that the conditional regression functions of the potential outcomes given the score are continuous at the cutoff. In other words, this continuity-based framework assumes that, at the cutoff point where the treatment status changes abruptly from control to treated, the average underlying features of the population only change smoothly, not abruptly. The abrupt change in treatment status induced by the sharp RD treatment assignment combined with this continuity condition allows researchers to recover the ATE at the cutoff—a quantity defined in terms of potential outcomes—from observed outcomes. This nonparametric (infinite population) identification result leads naturally to estimation and inference approaches that attempt to estimate the distance between two different and unknown conditional regression functions at the cutoff, using polynomials as approximation devices. Thus, this approach relies on nonparametric large-sample extrapolation methods.

The continuity-based framework regards the potential outcomes as random variables and the n observations as a random sample from a (super) population, an idea that we summarize in the following assumption:

Assumption 1. ***(Super population).*** *For $i = 1, 2, \ldots, n$: $(Y_i(0), Y_i(1), R_i)'$ is a random sample from a large (super) population.*

Moreover, it is common in this framework to also assume that each unit's potential outcome is only affected by that unit's treatment status—generalizations of which we discuss in detail below. This leads to each unit having exactly two potential outcomes, $Y_i(1)$ and $Y_i(0)$, where $Y_i(1)$ denotes the potential outcome when unit i receives treatment and $Y_i(0)$ denotes the potential outcome when unit i receives control. Thus, the observed outcome is

$$Y_i = Y_i(0) \cdot (1 - D_i) + Y_i(1) \cdot D_i = \begin{cases} Y_i(0) & if \quad D_i = 0 \\ Y_i(1) & if \quad D_i = 1 \end{cases}.$$

In our application, $Y_i(0)$ represents the child mortality rate in county i in the absence of Head Start assistance, while $Y_i(1)$ captures the same county's child mortality rate if the county receives Head Start assistance.

The main parameter of interest in this framework is the (super) population average response to treatment at the cutoff $\bar{r}$:

$$\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0) \,|\, R_i = \bar{r}].$$

Whether this is a causal parameter is a subject of some debate. On the one hand, this parameter is a function of the unit-level causal effect that captures the potential outcome difference between treated and untreated states, $Y_i(1) - Y_i(0)$, and in that sense may be regarded as causal. On the other hand, under the "no causation without manipulation" interpretation (Holland, 1986), and given that R_i is a continuous random variable, the probability of observing units at the cutoff is zero and thus this parameter cannot be directly conceived as an experiment that exogenously assigns treatment to a given population of units. This ambiguity does not occur in the local randomization framework we discuss below, where the parameter of interest is not the ATE at a point but within an interval, and the treatment assignment is directly conceived as an exogenous manipulation. Nevertheless, the parameter τ_{SRD} is often useful to test substantive theories in social, behavioral and biomedical sciences, as well as to develop policy recommendations.

We now discuss parametric and nonparametric identification of τ_{SRD}, and summarize the most common asymptotic inference methods based on those identification results. In essence, once identification is ensured, estimation and inference for τ_{SRD} involves modeling parametrically or approximating nonparametrically the two conditional regression functions $\mathbb{E}[Y_i(1)|R_i = r]$ and $\mathbb{E}[Y_i(0)|R_i = r]$ at (or near) the cutoff $\bar{r}$. These inference methods rely on particular large-sample Gaussian approximations to conduct estimation and inference. Typical regularity conditions include continuity of the running variable, and existence and boundedness of higher-order moments. We do not discuss this kind of technical regularity conditions in this paper, which may be found elsewhere in the literature (e.g., Calonico et al., 2016; Calonico, Cattaneo, & Titiunik, 2014b).

Parametric Estimation Methods

Ludwig and Miller (2007) employed parametric methods to estimate the ATE at the cutoff of Head Start assistance on child mortality. We classify as "flexible" parametric methods those that focus on observations in a neighborhood of the cutoff but (i) do not account for misspecification bias in estimation and inference procedures, and (ii) do not select this neighborhood using data-driven procedures based on nonparametric approximations.

The key assumption underlying flexible parametric identification, estimation, and inference is the following:

Assumption 2. ***(Parametric functions).*** *For some polynomial degree* $p = 0, 1, 2, \ldots,$

$$\mathbb{E}[Y_i(0) \,|\, R_i = r] = \beta_0^- + r\beta_1^- + \cdots + r^p\beta_p^-,$$

$$\mathbb{E}[Y_i(1) \,|\, R_i = r] = \beta_0^+ + r\beta_1^+ + \cdots + r^p\beta_p^+$$

for all $r \in [\bar{r} - h, \bar{r} + h]$*, where* h *is a positive, known bandwidth parameter.*

This assumption models the conditional expectations of the potential outcomes as parametric polynomial functions of the running variable, possibly locally to the cutoff. The choice of neighborhood $[\bar{r} - h,\ \bar{r} + h]$ is controlled by the choice of

bandwidth h. In this approach, however, Assumption 2 is taken as correct and thus no attention is paid to misspecification bias, and the choice of bandwidth h is typically ad hoc. The main estimation approach under Assumptions 1 and 2 (and regularity conditions) is parametric least-squares for a choice of polynomial order and neighborhood around the cutoff. We describe this procedure below.

Procedure 1 (Parametric Approach)

1. Select a neighborhood $W = [\bar{r} - h,\ \bar{r} + h]$, where h is the bandwidth, and the polynomial degree p. The bandwidth h is chosen by the researcher in an ad hoc way. In general, p is small ($p = 0$ or $p = 1$) when h is small, and large ($p = 4$ or $p = 5$) when h is large.
2. Drop all observations outside the neighborhood W, that is, keep only observations satisfying $\bar{r} - h \leq R_i \leq \bar{r} + h$.
3. Using least-squares regression, estimate the coefficients in the full treatment-interaction model.

$$Y_i = \alpha + \tau D_i + \beta_1 \bar{R}_i + \ldots + \beta_p \bar{R}_i^p + \zeta_1 \bar{R}_i D_i + \ldots + \zeta_p \bar{R}_i^p D_i + \varepsilon_i,$$

where $\bar{R}_i = R_i - \bar{r}$ is the re-centered running variable. The least-squares estimate of τ in the above regression model, denoted by $\hat{\tau}_{\text{SRD}}$, estimates τ_{SRD}. Heteroskedasticity-robust (or cluster-robust) standard errors for $\hat{\tau}_{\text{SRD}}$ are computed using standard least-squares algebra and are routinely calculated by statistical packages.

When the outcome of interest Y_i is replaced by some other pre-intervention regressors or unaffected post-treatment outcomes, the estimator $\hat{\tau}_{\text{SRD}}$ and associated inference procedures can be used to conduct the third falsification test described in the previous section, a "placebo test" that checks that a null effect is recovered for a variable that is, by construction, unaffected by the treatment. We present such results for our application in the Supporting Information Appendix.[5]

In empirical implementations, either h and p are small (local approximation) or h and p are large (global approximation), and both parameters are chosen ex ante by the researcher. Table III in Ludwig and Miller (2007, pp. 180–181) reports the RD treatment effects for the Head Start application using the flexible parametric approach in Procedure 1. Using the raw data, we re-estimate $\hat{\tau}_{\text{SRD}}$ and reproduce their results—see Table 3 for completeness and future comparability. While the point estimators are exactly equal, the standard errors differ slightly due to a change in degrees-of-freedom correction.

The results in Table 3 include local ($p = 1$) and global ($p = 4$) flexible parametric estimation for several neighborhoods around the cutoff: $[\bar{r} - 9,\ \bar{r} + 9]$ and $[\bar{r} - 18,\ \bar{r} + 18]$ for $p = 1$ and $[\bar{r} - 20,\ \bar{r} + 20]$ and the full data for $p = 4$. The latter global approach is not recommended in practice to estimate RD treatment effects because it (i) generates counterintuitive weighting schemes (Gelman & Imbens, 2014) and (ii) has erratic behavior near the cutoff (Runge's phenomenon in approximation theory). Global approximations are better suited to provide an overall, smooth graphical representation of RD design and falsification testing, as described in the previous section (see Calonico et al., 2015a, for more details). We include them here only for comparison and discussion.

The first panel in Table 3 shows the results for the main outcome in Ludwig and Miller (2007), mortality rates per 100,000 for children ages five to nine from causes

[5] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Table 3. Flexible parametric RD methods.

	Linear model ($p = 1$)		Quartic model ($p = 4$)	
	$h = 9$	$h = 18$	$h = 20$	Full sample
Ages 5–9, HS-targeted causes, post HS				
RD treatment effect	–1.895	–1.198	–2.751	–3.065
Parametric 95% CI	[–3.828, 0.038]	[–2.498, 0.101]	[–5.490, –0.012]	[–5.189, –0.940]
Parametric p-value	0.055	0.071	0.049	0.005
$N_W^- \mid N_W^+$	309 | 215	671 | 283	770 | 289	2489 | 294
Falsification tests, parametric *p*-values				
Ages 5–9, injuries, post-HS	0.953	0.318	0.933	0.887
Ages 5–9, HS-targeted, pre-HS	0.109	0.549	0.35	0.011

Notes: (i) All estimates are constructed using linear ordinary least-squares estimators with heteroskedasticity-robust standard errors; (ii) $N_W^- = \sum_{i=1}^n 1(\bar{r} - h \leq R_i < \bar{r})$, $N_W^+ = \sum_{i=1}^n 1(\bar{r} \leq R_i \leq \bar{r} + h)$.

affected by Head Start services. The linear specification yields a reduction of 1.895 points within ± 9 poverty index points around the cutoff, which then drops to 1.198 when the larger ± 18 neighborhood is considered. Both results are significant at 10 percent, with p-values of 0.055 and 0.071, respectively. The general direction of the effect is observed with a quartic polynomial, although the point estimates become somewhat larger in absolute value. As a placebo test, the bottom panel shows the p-values from RD effect estimation for two variables that should not have been affected by Head Start: mortality of children ages five to nine from injuries in the post-treatment period, and mortality of children ages five to nine from mortality causes targeted by Head Start (hereafter, HS-targeted causes) in the period before the program was adopted. With the only exception of the fully global specification (which is highly unreliable), the effects are statistically indistinguishable from zero at 10 percent level for both variables. Further falsification results are available in the Supporting Information Appendix.[6]

Nonparametric Local Polynomial Estimation Methods

An alternative to imposing a parametric form on the unknown regression functions is to leave these functions unspecified and employ modern nonparametric local polynomial methods for estimation and inference. Relative to the flexible parametric methods, a nonparametric local-polynomial approach has three distinctive features: (i) the bandwidth h is chosen in a data-driven way based on nonparametric approximations, (ii) the RD point estimator is asymptotically mean-squared-error (MSE) optimal, and (iii) inference procedures explicitly incorporate the effects of local parametric misspecification (i.e., nonparametric smoothing bias). For technical discussion on these points see Hahn et al. (2001), Imbens and Kalyanaraman (2012), Calonico et al. (2014b), and references therein.

The following assumption captures the essence of nonparametric identification, estimation, and inference methods for the RD treatment effect τ_{SRD}.

[6] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Assumption 3. ***(Nonparametric functions).*** $\mathbb{E}[Y_i(0)|R_i = r]$ *and* $\mathbb{E}[Y_i(1)|R_i = r]$ *are (at least) three times continuously differentiable at the RD cutoff* $r = \bar{r}$.

This assumption implies continuity of $\mathbb{E}[Y_i(0)|R_i = r]$ and $\mathbb{E}[Y_i(1)|R_i = r]$ at $r = \bar{r}$, the minimal requirement for nonparametric identification of τ_{SRD}. It also gives enough regularity to enable nonparametric estimation and inference of the RD treatment effect. This assumption is strictly weaker than Assumption 2, because the flexible parametric functional form imposed in Assumption 2 implies the smoothness of the conditional expectations imposed in Assumption 3.

From an implementation point of view, treatment effect estimation in the nonparametric approach is implemented in the same way as in the flexible parametric approach, the only difference being that the nonparametric approach often includes kernel weights that increase the relative weight of observations close to the cutoff. However, from a conceptual point of view, the nonparametric approach is fundamentally different from the flexible parametric approach when it comes to point estimation and inference. In the nonparametric approach, the only assumption is continuity (or differentiability) of the conditional expectations (Assumption 3); therefore, the least-squares estimation in Procedure 1 is misspecified by construction because, in general, Assumption 3 does not guarantee that the conditional expectation is exactly equal to the polynomial chosen for estimation. For this reason, the bandwidth h that determines the observations used via the neighborhood, $W = [\bar{r} - h,\ \bar{r} + h]$, must be chosen objectively to account for the presence of misspecification bias in estimation of and inference for the RD treatment effect. In this approach, h and p are always chosen to be small: local-constant ($p = 0$) or local-linear ($p = 1$), the latter being the preferred option in practice.

For point estimation, the most common approach is to choose the bandwidth h so that the resulting RD point estimator is approximately MSE-optimal. The key idea is that the bandwidth generates a trade-off between the bias and variance of the point estimator, and hence it can be chosen to optimally balance this trade-off. On the one hand, choosing a very small bandwidth reduces the bias because the regression approximation is better. However, the smaller the bandwidth, the smaller the number of observations used and hence the larger the variance. Conversely, a large bandwidth allows the researcher to use a larger number of observations, hence decreasing the variance, but the larger the bandwidth, the larger the bias, because the parametric polynomial approximation deteriorates as observations farther away from the cutoff are included. This trade-off is captured in the MSE, which can be written as the sum of the variance and the squared bias.

Specifically, under Assumptions 1 and 3 (and regularity conditions), and for a choice of polynomial approximation (controlled by the choice of p) and weights near the cutoff (controlled by the choice of kernel function $\mathcal{K}(\cdot)$), the MSE of the RD estimator can be approximated as:

$$\text{MSE}\,(\hat{\tau}_{\text{SRD}}) = \text{Bias}^2 + \text{Variance} \approx h^{2p+2}\text{B}^2 + \frac{1}{nh}\text{V},$$

where B and V denote constants that are specific to the kernel chosen and the data generating process. The above expression can easily be minimized, yielding the (infeasible) optimal bandwidth choice $h_{\text{MSE}} \propto (\text{V}/\text{B}^2)^{1/(2p+3)} n^{-1/(2p+3)}$. Employing h_{MSE} leads to a MSE-optimal RD estimator that uses only observations whose running variable falls within the neighborhood $W = [\bar{r} - h_{\text{MSE}},\ \bar{r} + h_{\text{MSE}}]$. In fact, this bandwidth choice makes the triangular kernel the optimal choice for weights (Cheng, Fan, & Marron, 1997). Using these ideas, Imbens and Kalyanaraman (2012) and Calonico et al. (2014b) recently developed, respectively, first and second generation plug-in bandwidth selectors for h_{MSE} (see Wand & Jones, 1995, for a review on

bandwidth selection methods). In our analysis, we employ the second-generation bandwidth selector proposed by Calonico et al. (2014b), denoted $\hat{h}_{\text{MSE}}$, which has better finite- and large-sample properties. See Cattaneo and Vazquez-Bare (2016) for a review of different bandwidth selection and related neighborhood selection methods used in RD designs.

For inference, it is essential to account for the bias in the approximation to the unknown regression functions. Nonparametric procedures are based on Assumption 3; this implies, by construction, that the regression functions near the cutoff will be misspecified in general. How much bias there is depends on the data generating process and bandwidth choice, with smaller bandwidths reducing bias at the expense of increasing variance. In particular, a well-known result is that the MSE-optimal bandwidth h_{MSE} is too "large" for the misspecification bias to be negligible in the distributional approximation of the estimator—implying that when the bandwidth chosen is h_{MSE}, inferences based on the Normal quantiles for the standard least-squares coefficients in Procedure 1 will be invalid. One alternative is to choose a bandwidth smaller than h_{MSE} (undersmoothing), but this approach is ad hoc, leads to power loss, and requires using different observations for point estimation and inference. An alternative, proposed by Calonico et al. (2014b), is to explicitly incorporate the bias in the distributional approximation, and construct a t-test statistic based on the ratio of the bias-corrected point estimator and a new variance estimator that takes into account the variability introduced in the bias-estimation step. Thus, the statistic employed re-centers and rescales the point estimator $\hat{\tau}_{\text{SRD}}$ to construct an inference procedure with demonstrably better theoretical and empirical properties (Calonico, Cattaneo, & Farrell, 2016, 2017). Using this approach, our inference results are captured by the robust p-value, which is simply the p-value calculated using the robust bias-corrected statistic.

The above discussion is summarized in the following procedure.

Procedure 2 (Nonparametric Approach)

1. Select a neighborhood $W = [\bar{r} - h,\ \bar{r} + h]$, with bandwidth h, polynomial degree p, and kernel weighting function $\mathcal{K}(\cdot)$. For the polynomial, usual choices are $p = 0$ (constant regression) or $p = 1$ (linear regression). For weights near the cutoff, the usual kernel choice is the triangular kernel, $\mathcal{K}(u) = 1 - |u|$. Finally, the bandwidth is chosen to be $h = \hat{h}_{\text{MSE}}$, that is, the MSE optimal bandwidth for the RD point estimator.
2. Drop all observations outside the neighborhood W, that is, keep only observations satisfying $\bar{r} - h \leq R_i \leq \bar{r} + h$.
3. Using weighted least-squares regression, estimate the coefficients in the full treatment-interaction model:

$$Y_i = \alpha + \tau D_i + \beta_1 \bar{R}_i + \ldots + \beta_p \bar{R}_i^p + \zeta_1 \bar{R}_i D_i + \ldots + \zeta_p \bar{R}_i^p D_i + \varepsilon_i$$

   with weights given by $\mathcal{K}(\bar{R}_i/h)$, and where $\bar{R}_i = R_i - \bar{r}$ is the re-centered running variable. The weighted least-squares estimate of τ in the above regression model, denoted by $\hat{\tau}_{\text{SRD}}$, estimates τ_{SRD}. Inference methods must account for the misspecification bias, and hence robust bias-corrected heteroskedasticity-robust (or cluster-robust) inference can be conducted using results in the literature (Calonico et al., 2016; Calonico et al., 2014b).

Table 4 presents the results from the nonparametric local polynomial analysis in the Head Start case study, estimated using the software packages described in Calonico et al. (2017). The table presents results from two analyses, one based on a local constant ($p = 0$) and the other on a local linear ($p = 1$) polynomial approximation.

Table 4. Robust bias-corrected local polynomial methods.

	Constant model ($p = 0$)		Linear model ($p = 1$)	
	$h = \widehat{h}_{MSE}$	$h = \widehat{h}_{FP1}$	$h = \widehat{h}_{MSE}$	$h = \widehat{h}_{FP1}$
Ages 5–9, HS-targeted causes, post-HS				
RD treatment effect	−2.114	−1.059	−2.409	−2.182
Robust 95% CI	[−4.963, −0.149]	[−4.34, −0.024]	[−5.462, −0.099]	[−5.722, −0.350]
Robust p-value	0.037	0.048	0.042	0.027
N_W^- | N_W^+	98 | 92	309 | 2015	234 | 180	309 | 215
h	3.235	9.000	6.810	9.000
Falsification Tests, robust *p*-values				
Ages 5–9, injuries, post-HS	0.880	0.960	0.728	0.787
Ages 5–9, HS-targeted, pre-HS	0.242	0.044	0.468	0.378

Notes: (i) Point estimators are constructed using local polynomial estimators with triangular kernel; (ii) "robust p-values" are constructed using bias-correction with robust standard errors as derived in Calonico, Cattaneo, and Titiunik (2014b); (iii) $\hat{h}_{MSE}$ corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico, Cattaneo, and Titiunik (2014b) and Calonico et al., (2016); (iv) $N_W^- = \sum_{i=1}^n 1(\bar{r} - h \leq R_i < \bar{r})$, $N_W^+ = \sum_{i=1}^n 1(\bar{r} \leq R_i \leq \bar{r} + h)$.

While it is often recommended to employ local linear RD estimators ($p = 1$), we also report the $p = 0$ case to facilitate later comparisons with the local randomization framework discussed below. In each case, the RD point estimate is estimated twice, once using the MSE-optimal bandwidth described above ($\hat{h}_{\text{MSE}}$) and the other using the bandwidth chosen by Ludwig and Miller (2007) in the flexible parametric approach ($\widehat{h}_{FP1}$). The local-polynomial results using the latter bandwidth are reported for comparability with the flexible parametric results reported in Table 3. Notice that $\widehat{h}_{FP1}$ appears to be indeed too "large" when compared to $\hat{h}_{\text{MSE}}$.

As in Table 3, the top panel reports results for the main outcome of interest—mortality of children ages five to nine from Head Start-targeted causes—and the bottom panel reports p-values for the two placebo outcomes—child mortality from non-HS-targeted causes and child mortality prior to implementation of Head Start. For the local constant polynomial model, the estimated MSE-optimal bandwidth $\hat{h}_{\text{MSE}}$ is 3.235 and the effect of the program based on this bandwidth is −2.114 (robust p-value 0.037). When the bandwidth is instead $\hat{h}_{\text{FP1}} = 9$, the point estimate associated with a local constant model decreases considerably in absolute value to −1.059. This point estimate, however, is likely to be considerably biased, since a bandwidth of nine is almost three times larger than the MSE-optimal choice, suggesting that a local constant approximation will fail to capture the curvature of the underlying regression function. For this reason, a local linear fit for this bandwidth is a preferable choice.

The last two columns of Table 4 show the results from a local linear model, both for the MSE-optimal bandwidth and the fixed bandwidth. As expected, the MSE-optimal bandwidth choice for the local linear model (6.811) is larger than the choice for the local constant model (3.235), since the former can accommodate observations further away from the cutoff and reduce the increased bias by allowing for a linear approximation (i.e., an approximation of higher order). The point estimator using this optimal bandwidth is −2.409, larger in absolute value but not too different from the local constant effect. And the point estimator using the fixed flexible parametric

bandwidth $\hat{h}_{\text{FP1}} = 9$ yields an effect of -2.182, now much closer to the results from the MSE-optimal bandwidth, as expected given that a slope is allowed. The effects on the placebo outcomes cannot be distinguished from zero at conventional levels if the preferred local linear approximation is used.

Additional Empirical Methods

Our discussion so far has focused exclusively on the most basic local polynomial approaches for estimation and inference in RD designs. These methods can be extended and complemented with more recent developments now available in the literature. We briefly discuss three related methods, which are also fully available in general-purpose software.

- *Different bandwidths for control and treatment units*. In some applications, it may be advisable to employ different bandwidths on the left and on the right of $\bar{r}$. Such bandwidths can also be chosen to be MSE-optimal, perhaps optimal in some other sense, or simply chosen in an ad hoc way. In the case of Head Start, we found that allowing for different bandwidths does not qualitatively affect the main empirical findings, though the left MSE-optimal bandwidth is substantially larger relative to the right MSE-optimal bandwidth, which is unsurprising given that there are many more observations on the left than on the right.
- *Coverage error optimal bandwidths for confidence intervals*. Calonico, Cattaneo, and Farrell (2016, 2017) recently developed an alternative way of choosing the bandwidth(s), which is tailored specifically to constructing confidence intervals with the smallest possible coverage error (as opposed to minimizing the MSE). This alternative bandwidth selector is smaller than h_{MSE} (in large samples), leading to fewer observations used for inference. For more discussion on bandwidth/window selection, see Cattaneo and Vazquez-Bare (2016). We report Head Start empirical results using these alternative bandwidths in the Supporting Information Appendix, where we find that results are consistent with those reported above.[7]
- *Pre-intervention covariate adjustments*. Calonico et al. (2016) develop formal local polynomial methods allowing for pre-intervention covariate adjustments. Applying these methods to the Head Start data gives qualitatively similar results to those reported herein, with smaller standard errors and a robust p-value that is reduced by almost 50 percent. Our companion replication files include these additional estimates for completeness.

Our companion replication software code implements all these additional empirical methods, which we do not report here to conserve space.

RD BASED ON RANDOMIZATION NEAR THE CUTOFF

An alternative approach to RD analysis assumes that, in a small neighborhood or window around the RD cutoff, the assignment of units to treatment or control status is random, as it would be in an experiment. This local randomization framework is conceptually different from the continuity-based framework described above. In the continuity-based framework, the parameter of interest is the difference between the average potential outcomes under treatment and control at the cutoff, and a

[7] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

central concern is whether these regression functions (whose functional form is fundamentally unknown) are well approximated at the cutoff. This concern about approximation/extrapolation is not as central in the local randomization framework because, in its simplest version, this approach implies that the regression functions are constant over all values of the running variable in a window around the cutoff. Instead, the central concern in the local randomization framework for RD analysis is the selection of the region or window where the treatment can be regarded as randomly assigned and, relatedly, whether the power of the inference procedures used is adequate (on the other hand, size and coverage can be well controlled if exact randomization-based inference methods are employed).

An advantage of the local randomization framework as developed by Cattaneo et al. (2015) is that—once the window where randomization holds is known or estimated—it justifies analyzing the RD design as one would analyze a randomized experiment. This means that there are two main alternatives for inference. The first alternative defines the ATE in the window as the parameter of interest and relies on large-sample approximations to derive the distribution of the relevant test statistics. This can be implemented by assuming either that the potential outcomes are random variables and relying on, say, large-sample approximations to the difference-in-means test-statistic, or by assuming that the potential outcomes are fixed and adopting a Neyman approach where the randomization distribution of the test statistic is approximated by letting the number of units in the experiment grow (Imbens & Rubin, 2015).

The second alternative assumes that the potential outcomes are non-random and the population of n units is fixed. This permits the use of finite-sample exact randomization inference methods, where the null distribution of the test statistic of interest is derived directly from the randomization distribution of the treatment assignment inside the window, leading to inferences that are exact in finite samples. This approach is sometimes called Fisherian inference, as it was first introduced by Fisher (1935). Fisherian randomization-based inference methods are an appealing alternative when there are few observations in the window where local randomization is plausible, which makes large-sample approximations unreliable. Since small sample sizes near the cutoff are common in RD applications, it is natural and advisable to employ Fisherian randomization-based inference methods to analyze RD designs under a local randomization framework. Of course, if the sample size inside the window where randomization is assumed to hold is large enough, then large-sample inference methods will also be appropriate (either Neyman-type or super-population type).[8]

In the remainder of this section, we discuss the local randomization approach to RD analysis adopting a Fisherian randomization-based framework for inference where both the sample and the potential outcomes are fixed. This framework was first developed for RD settings by Cattaneo et al. (2015). We extend their framework to develop inference methods that are appropriate not only when the potential outcomes are unaffected by the running variable, as assumed by Cattaneo et al. (2015), but also for the more general case where the running variable has a direct effect on the potential outcomes. We also introduce and discuss three new sensitivity procedures specifically tailored to RD designs. We apply all these methods to the Head Start empirical application, and we also report point estimates and related quantities (better justified in large samples) for comparison and completeness.

[8] An alternative approach is to use permutation-based inference, which employs random potential outcomes and different distributional assumptions/approximations under specific null hypotheses. For further discussion on the relationship between randomization inference and permutation inference methods, see Ernst (2004).

We start by modifying our notation to accommodate this local randomization approach to RD analysis based on Fisherian inference methods. We let $\mathbf{D}$ be the $n \times 1$ vector collecting the observed treatment assignment for the n observations, and the same for the score or running variable $\mathbf{R}$—that is, $\mathbf{D} = (D_1, D_2, ..., D_n)'$, and $\mathbf{R} = (R_1, R_2, ..., R_n)'$. Recall that throughout this paper, we use lower case to indicate fixed non-random variables and upper case to indicate random variables. The potential outcome of unit i for a given vector of treatment statuses $\mathbf{d}$ and a vector of scores $\mathbf{r}$ is $y_i(\mathbf{d}, \mathbf{r})$. The support of $\mathbf{D}$ is denoted by $\text{supp}(\mathbf{D}) := \mathcal{D} \subseteq \{0, 1\}^n$ and similarly $\text{supp}(\mathbf{R}) := \mathcal{R} \subseteq \mathbb{R}^n$. We collect these potential outcomes in a vector $\mathbf{y}(\mathbf{d}, \mathbf{r}) = (y_1(\mathbf{d}, \mathbf{r}), y_2(\mathbf{d}, \mathbf{r}), ..., y_n(\mathbf{d}, \mathbf{r}))'$. Finally, the observed outcome is $Y_i = y_i(\mathbf{D}, \mathbf{R})$, $i = 1, \ldots, n$, collected in the vector $\mathbf{Y} = \mathbf{y}(\mathbf{D}, \mathbf{R})$ taking values in the set $\mathcal{Y} \subseteq \mathbb{R}^n$. In principle, the potential outcomes are allowed to depend on the treatment assignments of all the units in the sample. Although the notation may seem complicated at first, we employ it to highlight that some inference procedures are robust to violations of some standard assumptions like SUTVA, as will be discussed shortly.

The notation does emphasize that in this sharp RD framework the only random quantity is $\mathbf{R}$ (and, by implication, $\mathbf{D}$, because of the RD treatment assignment rule). The potential outcomes could depend on the running variable in an unrestricted way. Previous work in this literature imposed the assumption $y_i(\mathbf{d}, \mathbf{r}) = y_i(\mathbf{d})$ near the cutoff, which implies that the average response to treatment is constant as a function of the running variable (Cattaneo et al., 2015). In the section below, we relax this requirement and propose instead a novel approach based on adjusting outcomes via flexible modeling before employing randomization inference methods. This approach is a strict generalization of the results available in the literature, in the sense that our methods reduce to those already available whenever unadjusted outcomes are used (i.e., whenever the adjustment is the identity function and hence no adjustment is done).

The local randomization framework for RD assumes that there is a window around the cutoff in which treatment is assigned as in a randomized experiment, that is, in which the assignment mechanism (or probability law) of $\mathbf{D}$ is completely known. We denote this window by $W_0 = [\bar{r} - w, \bar{r} + w]$, $w > 0$, which we assume symmetric only for simplicity, and let $\mathbf{R}_{W_0}$ be the subvector of $\mathbf{R}$ corresponding to the observations with $R_i \in W_0$—and similarly for other vectors like $\mathbf{D}_{W_0}$ and $\mathbf{Y}_{W_0}$. We also let $\mathcal{D}_{W_0}$ be the support of $\mathbf{D}_{W_0}$, and $\mathcal{R}_{W_0}$ be the support of $\mathbf{R}_{W_0}$. Finally, we let N_{W_0} denote the total number of units inside the window, $N^+_{W_0}$ the number of treated units within this window (i.e., units with $\bar{r} \leq R_i \leq \bar{r} + w$), and $N^-_{W_0}$ the number of control units within this window (i.e., $N^-_{W_0} = N_{W_0} - N^+_{W_0}$). As above, we continue to focus on a sharp RD, where all units with $R_i \geq \bar{r}$ receive treatment and all units with score below $\bar{r}$ receive the control condition, and therefore the distribution of the assignment vector is completely determined by the distribution of the running variable; the assignment probabilities can be characterized by modeling the distribution of R_i instead of that of D_i directly. Finally, as is standard in the analysis of randomized controlled trials using fixed potential outcomes (Imbens & Rubin, 2015; Rosenbaum, 2002b, 2010), our analysis proceeds conditionally on those units with R_i within the window W_0, where local randomization is assumed. Therefore, our analysis applies only to those units within the window, and does not necessarily apply more generally to other units that could have fallen within the window.[9]

We summarize our notion of local randomization in the following assumption:

[9] This is analogous to the analysis of experiments in the absence of random sampling, where the findings only apply to the specific sample available (internal validity) and not necessarily to the super-population to which the units belong (external validity).

Assumption 4. ***(Finite population and assignment mechanism).*** *There exists a window* $W_0 = [\bar{r} - w, \bar{r} + w]$, $w > 0$, *such that the following holds:*

1. *Non-random potential outcomes.* $\mathbf{y}(\mathbf{d}, \mathbf{r})$ *are fixed.*
2. *Unconfoundedness.* $\mathbb{P}(\mathbf{R}_{W_0} \leq \mathbf{r} \,;\, \mathbf{y}(\mathbf{d}, \mathbf{r})) = \mathbb{P}(\mathbf{R}_{W_0} \leq \mathbf{r})$*, for all vectors* $\mathbf{r} \in \mathcal{R}_{W_0}$.
3. *Known mechanism.* $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$ *is known for all vectors* $\mathbf{d} \in \mathcal{D}_{W_0}$.

Part 1 in Assumption 4 states that the potential outcomes are non-random—that is, they are fixed characteristics of the population of n units. Part 2 requires that, inside the window W_0 around the cutoff $\bar{r}$, the distribution of the vector of observed scores does not depend on the potential outcomes. Because in our framework $\mathbf{y}(\mathbf{d}, \mathbf{r})$ are fixed quantities, the assumption should be interpreted as implying that the distribution of $\mathbf{R}_{W_0}$ does not depend on the specific values that the (fixed) potential outcomes take. For further discussion of this assumption in the context of fixed potential outcomes, and its relationship to random potential outcomes, see Imbens and Rubin (2015, Section 3). This condition rules out any type of selection into treatment, as in a classical randomized experiment. Part 3 states that, inside the window, the distribution of the treatment vector is known to the researcher. In the Supporting Information Appendix, we offer a discussion linking this notion of local randomization using fixed potential outcomes to other notions of local randomization using random potential outcomes (building on Cattaneo et al., 2015; Sekhon & Titiunik, 2017) and to related ideas based on continuity-based identification, including Lee's (2008) model of imprecise manipulation.[10]

As is customary in the randomization inference literature, we consider two assignment mechanisms. The first one, usually known as fixed margins randomization, follows the distribution:

$$\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d}) = \frac{1}{\left|\mathcal{D}_{W_0}^{FM}\right|} = \binom{N_{W_0}}{N_{W_0}^{+}}^{-1}, \quad \forall \mathbf{d} \in \mathcal{D}_{W_0}^{FM},$$

where $\mathcal{D}_{W_0}^{FM} = \left\{\mathbf{d} \in \{0,1\}^{N_{W_0}} : \sum_{i=1}^{N_{W_0}} d_i = N_{W_0}^{+}\right\}$. In words, this assignment assumes that the number of treated units is fixed, and simply shuffles the treatment indicator across the sample. The second one, which will be referred to as Bernoulli trials, is characterized by:

$$\mathbb{P}(\mathcal{D}_{W_0} = \mathbf{d}) = q^{N_{W_0}^{+}}(1-q)^{N_{W_0}^{-}}, \quad \forall \mathbf{d} \in \mathcal{D}_{W_0}^{BE}$$

where $\mathcal{D}_{W_0}^{\mathrm{BE}} = \mathbf{d} \in \{0,1\}^{N_{W_0}}$ and q denotes the individual probability of treatment assignment. Intuitively, in this mechanism, treatment assignment is defined by simply flipping a coin for each unit in the sample. Note that in the Bernoulli assignment, unlike in the fixed margins assignment, the number of treated units is not fixed.

Both of these assignment mechanisms imply that the individual probabilities of being treated are the same for all units. Furthermore, we note that many different distributions on $\mathbf{R}_{W_0}$ can induce one of the above treatment assignment mechanisms, and for this reason we impose the assumption directly on $\mathbf{D}_{W_0}$ instead of on $\mathbf{R}_{W_0}$. Finally, in some applications, researchers may want to estimate q from the data using, for instance, its maximum likelihood estimator $\hat{q} = N_{W_0}^{+}/N_{W_0}$. In such cases, the associated randomization inference procedures may need adjustments to account for the additional estimation error introduced by $\hat{q}$, though at present

[10] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

no such adjustments are available in the literature. In the remainder of this article, we take fixed margins or complete randomization as the model for treatment assignment $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$ within W_0, which does not require estimating additional parameters, and we employ the Bernoulli assignment mechanism only for robustness checks.

Flexible Modeling of Outcomes Near the Cutoff

Without further assumptions, it is difficult to define a treatment effect of interest because each unit's potential outcome may depend not only on the unit's own score and treatment status, but on the other units' as well. In the most general case, the difference in potential outcomes for unit i under treatment and control status will not be a scalar parameter but a function of $(d_1, d_2, ..., d_{i-1}, d_{i+1}, ..., d_n)$ and $\mathbf{r}$. Moreover, the running variable will generally be continuous, and hence $\mathbf{r}$ can take uncountably many values. To make this problem more tractable, we assume that there is a transformation of the potential outcomes that removes the dependence on $\mathbf{r}$.

Assumption 5. ***(Transformed outcomes).*** *There exists a transformation $\phi(\cdot)$ such that, for all i with $R_i \in W_0$, the transformed potential outcomes only depend on d_{W_0}, that is,*

$$\phi\left(y_i\left(\mathbf{d}, \mathbf{r}\right), \mathbf{d}, \mathbf{r}\right) = \tilde{y}_i\left(\mathbf{d}_{W_0}\right) \quad \forall\, \mathbf{r} \in \mathcal{R}.$$

Assumption 5 reduces the number of potential outcomes of interest for each individual to a finite number, since it makes it only a function of $\mathbf{d}_{W_0}$, which takes (at most) $2^{N_{W_0}}$ values. The simplest and perhaps most natural way to satisfy this assumption is to assume an exclusion restriction requiring that, inside W_0, the potential outcomes depend on the score only through the value of the treatment indicator—that is, only on whether R_i is smaller or greater than the cutoff $\bar{r}$, but not on the particular value of R_i. This is precisely what Cattaneo et al. (2015) assumed in their framework. Assumption 5 above weakens this requirement by allowing R_i to affect the potential outcome of individual i not only through D_i, but also directly.

To implement this approach, we propose a model that simultaneously allows us to make the dependence of the potential outcomes on $\mathbf{r}$ more tractable, and link the local randomization RD approach to the continuity-based framework discussed above. The particular model we consider is as follows:

$$y_i\left(\mathbf{d}, \mathbf{r}\right) = \begin{cases} \alpha_i\left(\mathbf{d}_{W_0}\right) + \left(r_i - \bar{r}\right)\beta_1^- + \left(r_i - \bar{r}\right)^2\beta_2^- + \cdots + \left(r_i - \bar{r}\right)^p\beta_p^- & \text{if } d_i = 0 \\ \alpha_i\left(\mathbf{d}_{W_0}\right) + \left(r_i - \bar{r}\right)\beta_1^+ + \left(r_i - \bar{r}\right)^2\beta_2^+ + \cdots + \left(r_i - \bar{r}\right)^p\beta_p^+ & \text{if } d_i = 1 \end{cases}$$

for all i such that $R_i \in W_0$ and p is a non-negative integer denoting the degree of the polynomial.

First, we note that this specification assumes that each unit's potential outcome depends on the treatment assignment vector $\mathbf{d}$ and on the unit's score r_i, but not on other units' score values. Second, the model states that the direct effect of the score on the potential outcomes can be captured by a polynomial of order p on the unit's score, with slopes that are constant for all individuals on the same side of the cutoff—that is, for all units with the same treatment status. Third, our model allows the intercept (captured by the term $\alpha_i(\mathbf{d}_{W_0})$) to vary arbitrarily by unit; these intercepts capture the effect of treatment net of the score's "direct effect." Thus, adopting a model where the potential outcomes are directly affected by the value of the score through a polynomial, we can directly connect the local randomization approach with the continuity-based framework described above if the unknown conditional

regression functions of the potential outcomes were given (or approximated) by two polynomial functions on either side of the cutoff.

Finally, a particularly important case of this parametric model of transformed (potential) outcomes occurs when $\beta_{\mathrm{j}}^{-} = 0 = \beta_{\mathrm{j}}^{+}$ for $j = 1, 2, \cdots, p$, which leads to

$$y_i(\mathbf{d}, \mathbf{r}) = \alpha_i(\mathbf{d}_{W_0}) \equiv \tilde{y}_i(\mathbf{d}_{W_0})$$

for all units with their score inside W_0—the exclusion restriction adopted by Cattaneo et al. (2015). When this restriction holds, the specific value of R_i does not affect the potential outcomes directly. This leads to potential outcomes functions that are constant functions of the score inside W_0, removing all uncertainty regarding functional form. The more general polynomial model relaxes this assumption by allowing the score to affect the potential outcomes directly—albeit by making a very specific functional form assumption.

Given the general polynomial model, the transformed potential outcomes are defined as:

$$\tilde{y}_i(\mathbf{d}_{W_0}) := \begin{cases} y_i(\mathbf{d}, \mathbf{r}) - (r_i - \bar{r})\beta_1^{-} - \cdots - (r_i - \bar{r})^p \beta_p^{-} = \alpha_i(\mathbf{d}_{W_0}) & \text{if } d_i = 0 \\ y_i(\mathbf{d}, \mathbf{r}) - (r_i - \bar{r})\beta_1^{+} - \cdots - (r_i - \bar{r})^p \beta_p^{+} = \alpha_i(\mathbf{d}_{W_0}) & \text{if } d_i = 1 \end{cases}$$

for all i such that $R_i \in W_0$. Thus, the transformed potential outcomes isolate the portion of the potential outcome that is related to the treatment but unrelated to the particular value taken by the running variable.

Since the values of $\beta_1^{-}, \ldots, \beta_{\mathrm{p}}^{-}, \beta_1^{+}, \ldots, \beta_{\mathrm{p}}^{+}$ are unknown, we must calculate them. Given our model, this is easily done using least-squares methods, where two separate regressions of the observed outcome on a polynomial of order p on the score inside the window W_0 are run, separately for observations above and below the cutoff. We let $\tilde{\beta}_{\mathrm{j}}^{-}$ and $\tilde{\beta}_{\mathrm{j}}^{+}$ denote the values of the slopes calculated by least-squares methods for $k = 1, 2, \ldots, p$. This type of adjustment does not come from a model for a random population but rather from an algorithmic fit, since the potential outcomes are non-stochastic (see, e.g., Rosenbaum, 2002a). Given these definitions, the observed transformed outcomes can be calculated as $\tilde{Y}_i(\mathbf{D}_{W_0}) = y_i(\mathbf{D}, \mathbf{R}) - (R_i - \bar{r})\tilde{\beta}_1^{-} - (R_i - \bar{r})^2\tilde{\beta}_2^{-} - \cdots - (R_i - \bar{r})^p\tilde{\beta}_p^{-}$ for units in the window to the left of the cutoff and $\tilde{Y}_i(\mathbf{D}_{W_0}) = y_i(\mathbf{D}, \mathbf{R}) - (R_i - \bar{r})\tilde{\beta}_1^{+} - (R_i - \bar{r})^2\tilde{\beta}_2^{+} - \cdots - (R_i - \bar{r})^p\tilde{\beta}_p^{+}$ for units in the window to the right of the cutoff; the vector that collects them is denoted by $\tilde{\mathbf{Y}}(\mathbf{D}_{W_0})$. It is important to emphasize, however, that this approach requires a correct specification of the polynomial model to yield valid results. Just as in the parametric modeling scenario discussed earlier, this adjustment approach ignores the possibility of misspecification, so using an incorrect model will generally lead to invalid results, with the difference that in this case the model needs to correctly fit the data not in a population but in the observed sample.

Testing the Sharp Null of No Effect

Having stated a randomization mechanism and a transformation model for the potential outcomes as in Assumptions 4 and 5, the null hypothesis that the treatment has no effect on any unit inside the window W_0 where local randomization holds can be tested using Fisherian randomization-based inference. This hypothesis is usually known as the *sharp null hypothesis of no effect*, and in our context is defined as $\mathrm{H}_0 : \alpha_i(\mathbf{d}_{W_0}) = \alpha_i(\mathbf{d}^*_{W_0})$ for any $\mathbf{d}_{W_0}, \mathbf{d}^*_{W_0}$ and for all i such that $R_i \in W_0$. In words, this means that under the null hypothesis, the (adjusted) potential outcomes are the same regardless of the treatment assignment. Letting this common value be $\alpha_i(0)$, under this hypothesis, we have $\tilde{y}_i(\mathbf{d}_{W_0}) = \tilde{y}_i(\mathbf{d}^*_{W_0}) = \alpha_i(0)$ for any $\mathbf{d}_{W_0}, \mathbf{d}^*_{W_0}$ and all i

such that $R_i \in W_0$. Collecting all observed transformed outcomes in W_0 in the vector $\tilde{\mathbf{Y}}(\mathbf{D}_{W_0})$ and all $\alpha_i(0)$ in W_0 in the vector $\boldsymbol{\alpha}^0_{W_0}$, under H_0 we have that $\tilde{\mathbf{Y}}_{W_0} = \boldsymbol{\alpha}^0_{W_0}$. But $\boldsymbol{\alpha}^0_{W_0}$ is constant under all realizations of $\mathbf{D}$. Therefore, any test-statistic $T(\mathbf{D}_{W_0}, \tilde{\mathbf{Y}}_{W_0})$ satisfies $T(\mathbf{D}_{W_0}, \tilde{\mathbf{Y}}_{W_0}) = T(\mathbf{D}_{W_0}, \boldsymbol{\alpha}^0_{W_0})$ under H_0, implying that its null distribution is known—because the only source of randomness in $T(\mathbf{D}_{W_0}, \boldsymbol{\alpha}^0_{W_0})$ is the treatment assignment mechanism, whose distribution is assumed to be known.

It follows that, under the sharp null hypothesis, we can compute an exact p-value for any observed value T_{obs} of the test-statistic, as described in the following procedure.

Procedure 3 (Local Randomization Approach)

1. Select a window $W_0 = [\bar{r} - w, \bar{r} + w]$ where local randomization is assumed to hold, and a model of potential outcomes adjustment. Typical models include constant regression ($p = 0$, no adjustment) and linear regression ($p = 1$). This gives the (adjusted) outcomes $(\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_{N_W})$.
2. Assume a treatment assignment mechanism. We employ fixed margins or complete randomization by default, but also use Bernoulli trials for robustness checks.
3. Select a test statistic. Usual examples include difference-in-means, Kolmogorov-Smirnov, and Wilcoxon-Mann-Whitney statistics. The finite-sample exact p-value for the null hypothesis of no treatment effect is calculated using randomization-inference via permutation of the treatment status of units under the sharp null hypothesis. See the Supporting Information Appendix for details on numerical implementation.[11]

Randomization inference methods can be extended to any sharp null hypothesis, that is, any null hypothesis under which the missing potential outcomes can be imputed. Thus, for example, it can be used to test that the treatment effect is constant and additive under no interference, or to test for interference after imposing a parametric model as in Bowers, Fredrickson, and Panagopoulos (2013).

A common choice for the statistic T is the difference in means. Letting $\mathcal{I}_0 = \{i : R_i \in W_0\}$ be the set of units inside the window, this statistic is

$$T_{\text{DM}} = \frac{\sum_{i \in \mathcal{I}_0} \tilde{Y}_i D_i}{\sum_{i \in \mathcal{I}_0} D_i} - \frac{\sum_{i \in \mathcal{I}_0} \tilde{Y}_i (1 - D_i)}{\sum_{i \in \mathcal{I}_0} (1 - D_i)},$$

or some variation of it such as its absolute value or the Studentized version that divides by the standard error. Another choice is to use rank-based statistics such as the Wilcoxon or Mann-Whitney statistics, or the Kolmogorov-Smirnov (KS) statistic, $T_{\text{KS}} = \sup_y |\hat{F}_{D=1}(y) - \hat{F}_{D=0}(y)|$, where $\hat{F}_{D=1}(y)$ and $\hat{F}_{D=0}(y)$ are estimates of the distribution function of the (transformed) outcome for the treated and control units, respectively. See Lehmann (1998), Rosenbaum (2002b), and Imbens and Rubin (2015) for discussions on the choice of statistic and power for randomization tests.

Naturally, the application of Fisherian inference to RD analysis is straightforward if the window W_0 around the cutoff where local randomization holds is known to the researcher. In practice, however, this window will be unknown and must be estimated. We choose this window following the procedure proposed in Cattaneo et al.

[11] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

(2015), who suggest finding an interval around the cutoff in which pretreatment covariates are balanced between treated and control units. We briefly discuss implementation issues related to the choice of W_0 below when we present our empirical illustration.

Once the observations satisfying Assumptions 4 and 5 are identified (i.e., once a window around the RD cutoff is selected and, if needed, a model of potential outcomes is imposed), the distribution of any test statistic is known under the sharp null hypothesis of no effect because the only randomness in the model is generated by the known assignment mechanism. Therefore, once a test statistic is selected and an assignment mechanism is specified, the sharp null of no effect can be tested by computing the exact p-value $\mathbb{P}(T(\mathbf{D}_{W_0}, \boldsymbol{\alpha}^0_{W_0}) \geq T_{obs})$. While conceptually straightforward, computing this probability exactly is usually impossible in practice because its computation requires calculating the test statistic under all possible configurations of treatment assignments, $\mathbf{D}_{W_0}$, as specified by the assignment mechanism, $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$. The solution to this numerical problem is to approximate the p-value by simply simulating different treatment assignments $\mathbf{D}_{W_0}$, according to $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$, and then computing the corresponding p-value using the simulated statistics. To conserve space, we outline in the Supporting Information Appendix how this procedure is carried out in practice, but we highlight that this method is in fact already implemented in our companion Stata and R software implementations.[12]

Estimands, Estimation, and Inference under SUTVA

Testing the sharp null hypothesis does not require any assumptions beyond having (transformed) potential outcomes that depend only on $\mathbf{D}$ and knowledge of the randomization distribution of $\mathbf{D}$. However, point and confidence interval estimation, as well as testing of other hypotheses, requires further assumptions. A very common simplifying assumption, and one that was adopted by the continuity-based approach discussed above, is to assume that units do not interfere with each other, usually called the *stable unit treatment value assumption* (SUTVA).

Assumption 6. ***(Local SUTVA).*** *For all i with $R_i \in W_0$, $\tilde{y}_i(\mathbf{d}_{W_0}) = \tilde{y}_i(d_i)$.*

Assumption 6 simply states that the potential outcomes for unit i only depend on unit i's treatment assignment. In other words, under SUTVA, each unit has only two transformed potential outcomes that simplify to $\tilde{y}_i(1) = y_i(1, r) - \beta_1^+(r - \bar{r}) - \beta_2^+(r - \bar{r})^2 - \cdots - \beta_{\mathrm{p}}^+(r - \bar{r})^p = \alpha_i(1)$ and $\tilde{y}_i(0) = y_i(0, r) - \beta_1^-(r - \bar{r}) - \beta_2^-(r - \bar{r})^2 - \cdots + \beta_{\mathrm{p}}^-(r - \bar{r})^p = \alpha_i(0)$. Under this assumption, the observed unadjusted and transformed potential outcomes can be written, respectively, as $Y_i(D_i, R_i) = D_i y_i(D_i, R_i) + (1 - D_i) y_i(1 - D_i, R_i)$ and $\tilde{Y}_i(D_i) = D_i \tilde{y}_i(D_i) + (1 - D_i)\tilde{y}_i(1 - D_i)$, for $D_i \in \{0, 1\}$, and we can define the treatment effect for unit i as $\tau_i := \tilde{y}_i(1) - \tilde{y}_i(0) = \alpha_i(1) - \alpha_i(0)$. In principle, the treatment effect can vary arbitrarily across units. A common parameter of interest in this case is the ATE for units with scores in W_0 that can be defined as $\tau_{W_0} = \frac{1}{N_{W_0}} \sum_{i \in \mathcal{I}_0} (\tilde{y}_i(1) - \tilde{y}_i(0))$ where, as before, $\mathcal{I}_0$ is the set of units inside the window.

When the transformed outcomes come from a least-squares fit at each side of the cutoff, τ_{W_0} captures the difference in the intercepts of the two regression functions.

[12] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

In other words, we have $y_i(d, \bar{r}) = \alpha_i(d)$ for $d \in \{0, 1\}$. In this case, and given our definition of the transformed outcomes, it follows that

$$\tau_{W_0} = \frac{1}{N_{W_0}} \sum_{i \in \mathcal{I}_0} (\tilde{y}_i(1) - \tilde{y}_i(0)) = \frac{1}{N_{W_0}} \sum_{i \in \mathcal{I}_0} (\alpha_i(1) - \alpha_i(0)),$$

which is the ATE when the running variable is evaluated at the cutoff $\bar{r}$. This is the closest parameter to the usual RD estimand in the continuity-based framework. Note, however, that this parameter is defined relative to the sample determined by the choice of window W_0 where local randomization is assumed to hold (or, at least, assumed to give a good approximation). Therefore, our results only apply to this sample, and do not necessarily generalize to other possible samples that would be realized in a hypothetical repeated sampling setting (recall that **R** is assumed random in our framework). A complete discussion connecting "population" and "super-population" analysis, in the context of classical randomized experiments and related settings, is given in Imbens and Rubin (2015) and references therein.

A possible estimator for the ATE is the difference in means of the observed transformed outcomes for treated and control units inside the window,

$$\hat{\tau}_{W_0} = \frac{1}{N^+_{W_0}} \sum_{i \in \mathcal{I}_0} \tilde{Y}_i D_i - \frac{1}{N^-_{W_0}} \sum_{i \in \mathcal{I}_0} \tilde{Y}_i (1 - D_i).$$

When the transformed outcomes are obtained by subtracting the slopes from a linear or polynomial regression of the outcome on the score at each side of the cutoff, $\hat{\tau}_{W_0}$ corresponds to the difference in the intercepts of the two regressions. Hence, the estimator coincides with the local linear regression estimator using a uniform kernel and bandwidth equal to half the length of the window W_0 (i.e., $h = w$), one of the most popular RD estimators in the continuity-based approach.

The difference-in-means statistic is appealing because it can be used directly to obtain a point estimator of the ATE, while for other statistics one needs to construct a Hodges-Lehmann estimate (under additional assumptions). Moreover, whenever transforming the outcomes is not needed, this statistic is unbiased for the ATE when the treatment assignment follows a fixed-margin randomization scheme or Bernoulli trials, conditionally on the observed sample. The main drawback, however, is that when arbitrary heterogeneity of treatment effects is allowed, randomization methods can no longer be applied to perform inference for the ATE because, unlike the sharp null hypothesis of no effect, the null hypothesis of no ATE is not sharp—that is, it does not allow the researcher to impute all the missing potential outcomes. Therefore, the statistical significance of the estimated ATE must be based on a large-sample approximation to its distribution. In the finite population framework, this leads to a Neyman approach to causal inference, which relies on a Normal distributional approximation with a two-sample standard error. This inference approach is asymptotically conservative.

There are, however, some restrictions that can be imposed to the treatment effects under which the sharp null of no effect is informative about the ATE. For example, if the treatment effect is non-negative—$y_i(1) \geq y_i(0)$ for all observations in the window—then the ATE is equal to zero if and only if $y_i(1) = y_i(0)$ for all i in the window, that is, if the sharp null of no effect holds. Hence, rejecting the sharp null implies rejecting the null that the ATE is equal to zero.

Randomization-Based Analysis of Head Start

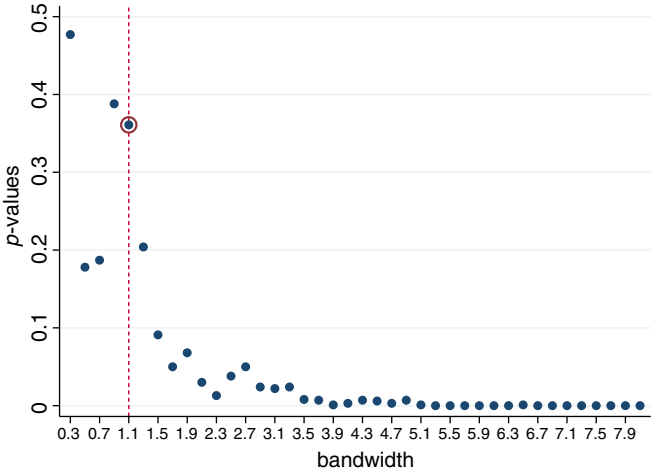
We now apply the Fisherian framework to perform local randomization RD analysis in the case of Head Start. We start by choosing the window in which the local randomization assumption is plausible based on pretreatment covariates. The window selection procedure treats each pretreatment covariate as an outcome, and tests the sharp null hypothesis of no effect in windows of increasing length. The chosen window is one of the largest windows such that the minimum p-value across all tests and for all (or most) smaller windows is above a certain cutoff (Cattaneo et al., 2015). We emphasize that multiple testing adjustment is not required in this context, and possibly not advisable either, since the goal is to be conservative and reject the null hypothesis with high probability to ensure that the selected window is plausibly consistent with the (local) randomization assumption. In other words, adjusting p-values for multiple testing will necessarily lead to a larger selected window, and hence we prefer to be (overly) conservative when it comes to window selection. For implementation, our chosen test statistic is the KS statistic, and our covariates come from the 1960 census, including population, schooling, and demographic characteristics. The complete list of variables, together with the results for other choices of statistics and related methods, are in the Supporting Information Appendix.[13]

We start with a window of 0.3 and increase it by 0.2 at each step. Panel (a) in Figure 2 depicts the minimum p-value from a KS test as a function of window length. Since we re-center the running variable, $\bar{R}_i = R_i - 59.1984$, the cutoff is $\bar{r} = 0$. As shown in Figure 2a, although the sequence of p-values is not monotonic, it stabilizes below 0.2 after a window length of 2×1.1. Our chosen window is therefore $\hat{W}_0 = [-\hat{w}, \hat{w}] = [-1.1, 1.1]$, which is about a third of the MSE-optimal bandwidth ($\hat{h}_{\text{MSE}} = 3.235$), and considerably smaller than the ones used in the flexible parametric specifications ($\hat{h}_{\text{FP1}} = 9$ and $\hat{h}_{\text{FP2}} = 18$). We explore the robustness of our results to the choice of window below, and in the Supporting Information Appendix.[14]

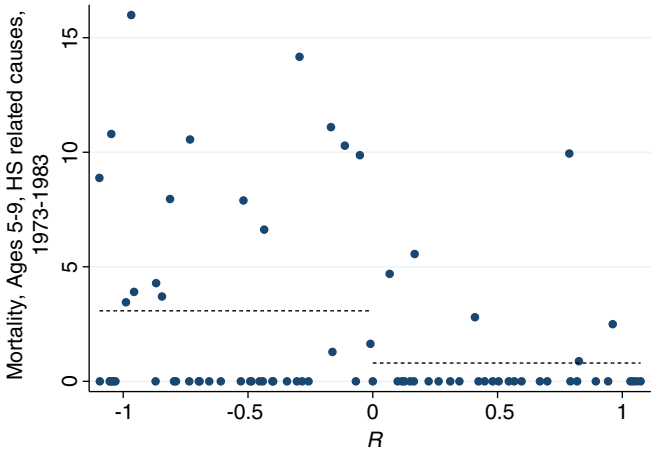
Panel (b) in Figure 2 shows a scatter plot of the outcome of interest, county-level mortality rate of children age five to nine from HS-targeted causes, against the re-centered running variable in the selected window. This window includes 43 and 33 observations below and above the cutoff, respectively. Even though the control and treated groups seem slightly unbalanced in terms of size, a binomial test that the probability of being treated is 0.5 does not reject the null hypothesis (p-value = 0.302). Moreover, consistent with the idea of local randomization, the scatter plot in Figure 2b suggests no clear relationship between the running variable and the outcome.

The point estimates—calculated as differences in means and shown as horizontal lines in Figure 2b—and their corresponding p-values can be seen in Table 5. For comparison, we also show the results for windows of length equal to (twice) the MSE-optimal bandwidth and the bandwidth chosen for the flexible parametric specification. The results are consistent with the findings from previous sections. In the case of no adjustment/transformation ($p = 0$, and SUTVA)—that is, assuming $\tilde{y}_i(\mathbf{d}_{W_0}) = y_i(\mathbf{d}, \mathbf{r}) = y_i(d_i)$—the point estimate in $\hat{W}_0 = [-1.1, 1, 1]$ is very close in magnitude and with the same sign and statistical significance as the ones found in previous sections, with a p-value below 0.01. Notice that this p-value does not require imposing SUTVA (or a treatment effect model), but it does

[13] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.
[14] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.



(a) *p*-values and Window Length



(b) Child Mortality and Poverty Index

Notes: (i) Panel (a) shows the minimum *p*-value from a Kolmogorov-Smirnov test for all covariates and for each window (selected window is [−1.1, 1.1]); (ii) panel (b) shows the scatter plot of the outcome of interest against the re-centered running variable $\bar{R}_i = R_i - 59.1984$ in the selected window (dotted lines represent the average outcome below and above the cutoff).

Figure 2. Window selection and outcome of interest.

require the exclusion restriction implied by Assumption 5. On the other hand, the estimated effect is closer to zero and not significant for the larger windows $\hat{W}_{\text{MSE}} = [-\hat{h}_{\text{MSE}}, \hat{h}_{\text{MSE}}] = [-3.235, 3.235]$ and $\hat{W}_{\text{FP}} = [-\hat{h}_{\text{FP1}}, \hat{h}_{\text{FP1}}] = [-9, 9]$.

When transforming the outcomes using a linear model ($p = 1$), the results for the outcome of interest using $\hat{W}_0$ are preserved—the point estimate increases slightly in absolute value and remains strongly statistically significant. The estimates for the two larger windows increase considerably in absolute value relative to the $p = 0$ results and become statistically significant, now yielding similar conclusions to the results calculated within $\hat{W}_0$. Regarding the falsification tests, the effects on both placebo outcomes are indistinguishable from zero in $\hat{W}_0$ for both the constant and linear specifications. For the wider windows, the results are mixed, with the

Table 5. Local randomization methods.

	No transformation ($p = 0$)			Linear transformation ($p = 1$)		
	$W = \hat{W}_0$	$W = \hat{W}_{MSE}$	$W = \hat{W}_{FP}$	$W = \hat{W}_0$	$W = \hat{W}_{MSE}$	$W = \hat{W}_{FP}$
Ages 5–9, HS-targeted causes, post-HS						
RD treatment effect	−2.280	−1.240	−0.691	−2.515	−3.726	−1.895
Fisher's p-value	0.009	0.156	0.145	0.006	0.000	0.000
N_W^- | N_W^+	43 | 33	98 | 92	309 | 215	43 | 33	98 | 92	309 | 215
w	1.100	3.235	9.000	1.100	3.235	9.000
Falsification tests, Fisher's p-values						
Ages 5–9, injuries, post-HS	0.699	0.606	0.059	0.185	0.762	0.908
Ages 5–9, HS-targeted, pre-HS	0.937	0.012	0.731	0.227	0.777	0.013

Notes: (i) Point estimators are constructed using difference-in-means of untransformed and transformed outcomes, respectively, with a uniform kernel; (ii) randomization p-values are obtained using 10,000 permutations; (iii) w corresponds to length of the half windows around zero of the centered running variable ($\bar{R}_i = R_i - \bar{r}$), that is, $W = [-w, w]$; (iii) $\hat{W}_0 = [-\hat{w}, \hat{w}]$ is selected as discussed in the text, employing the method described in Cattaneo et al. (2015), while $\hat{W}_{MSE} = [-\hat{h}_{MSE}, \hat{h}_{MSE}]$ and $\hat{W}_{FP} = [-\hat{h}_{FP}, \hat{h}_{FP}]$; (iv) $N_W^- = \sum_{i=1}^n 1(\bar{r} - w \leq R_i < \bar{r})$, $N_W^+ = \sum_{i=1}^n 1(\bar{r} \leq R_i \leq \bar{r} + w)$.

linear adjustment rejecting the null for one of the placebo outcomes at 1 percent in $\hat{W}_{\text{FP}}$. This is not surprising because for the larger windows the assumption of local randomization is implausible.

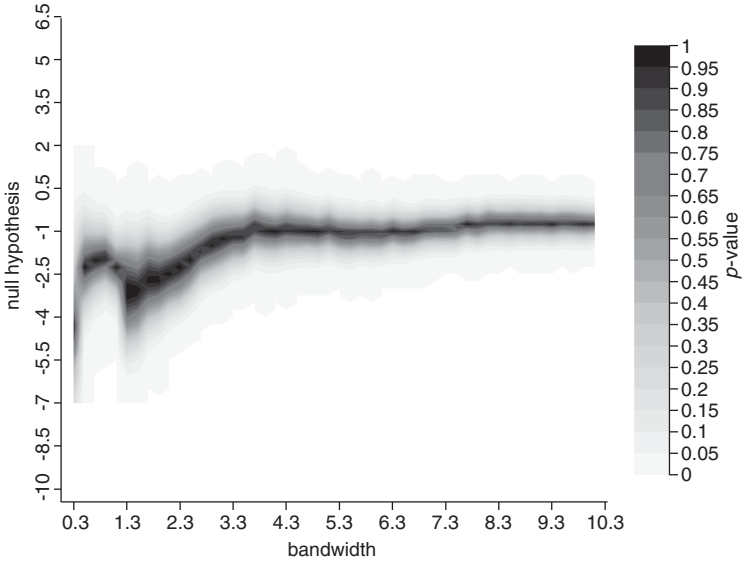
Finally, under our assumptions (and SUTVA), a constant treatment effect model of the form $\alpha_i(1) = \alpha_i(0) + \tau$ can be used together with the randomization-based procedures to calculate 95 percent confidence intervals, by collecting the range of τ_0 values that fail to be rejected in a test of the null hypothesis $\text{H}_0 : \alpha_i(1) = \alpha_i(0) + \tau_0$. This hypothesis is equivalent to testing the sharp null hypothesis on the adjusted transformed potential outcomes, where the adjustment removes the hypothesized treatment effect from the treated transformed outcomes, that is, $\tilde{\tilde{Y}}_i = \tilde{Y}_i - D_i\tau_0$ and $T(\mathbf{D}_{W_0}, \tilde{\tilde{\mathbf{Y}}}_{W_0}) = T(\mathbf{D}_{W_0}, \boldsymbol{\alpha}^0_{W_0})$ under H_0. We report these randomization-based confidence intervals in the next section.

Sensitivity Analysis and Robustness Checks

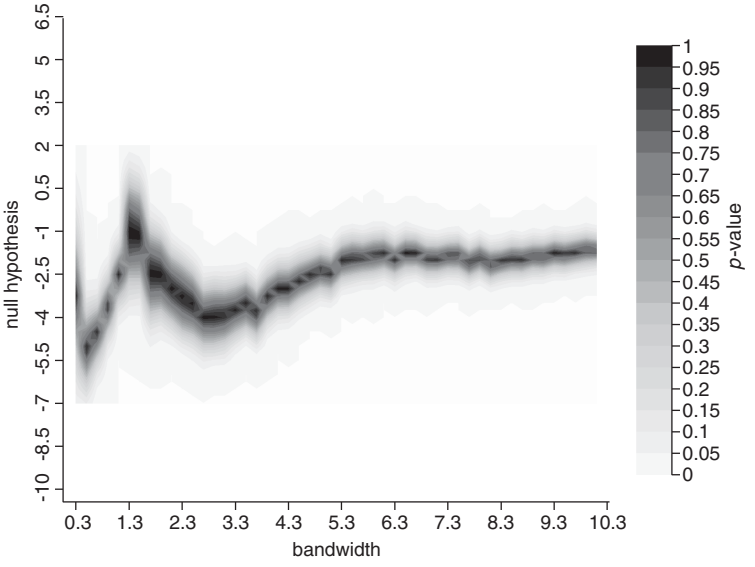
We now develop three novel methods to assess the robustness of the results under the local randomization assumption, specifically tailored to the particular features of RD designs. We focus on three main underlying assumptions or choices: (i) window length, (ii) interference between units, and (iii) misspecification of the randomization mechanism.

Sensitivity to Window Length

A natural question when implementing the local randomization framework for RD analysis is how the results vary for different window choices. To assess the sensitivity of the inferential results to this choice, Figure 3 displays the randomization p-values for a list of window choices and a grid of values for the treatment effect under an additive treatment effect model using constant (left) and linear (right) adjustments. The black region indicates p-values close to one, so for each window length, projecting this region onto the vertical axis gives the range of treatment effect values that cannot be rejected in a hypothesis test that assumes a constant



(a) $p = 0$ (Constant Model)



(b) $p = 1$ (Linear Model)

Notes: (i) Sensitivity to window length (x-axis) using constant additive treatment effect model (y-axis); (ii) randomization p-values are obtained using 1,000 permutation; (iii) bandwidth is w and corresponds to length of the half windows around zero of the centered running variable $\bar{R}_i = R_i - \bar{r}$, that is, $W = [-w; w]$.

Figure 3. Sensitivity of p-values to window length choice.

treatment effect model. In other words, this region gives a confidence interval for the treatment effect via inversion methods.[15]

[15] More precisely, the figure is constructed in the following way. First, define a grid of values for the treatment effect and a set of window lengths. Then, for each window and hypothesized treatment effect

Figure 3 suggests that the results are robust to the window length, starting at values of the treatment effect around −2.5 and then stabilizing around −1.3. Note, however, that as the window length increases the estimates stop being significant, as the 95 percent confidence interval—the range of treatment effect values that are not rejected with a 5 percent-level test—includes zero. This is consistent with our findings in Table 5. The graph with linear adjustment shows the same qualitative results although with more variability, which is understandable given that for small window lengths the fit can be sensitive to outliers. In all, while the treatment effect vanishes for large windows, in the range of bandwidths for which the local randomization assumption is plausible, the evidence for a significant negative treatment effect seems robust. In the Supporting Information Appendix we present formal empirical results for $w \in \{0.9, 1.1, 1.3, 1.5, 2.7\}$.

Inference in Presence of Interference

We now consider sensitivity of the results to the SUTVA assumption. While this is a useful and extremely common assumption, there are some scenarios under which it might be violated in RD applications. For instance, greater vaccination rates among children in treated counties may lower disease contagion in geographically adjacent counties, even if adjacent counties do not receive the program. In this case, the potential outcome of a county would depend not only on the county's own treatment status, but also on the treatment status of adjacent counties. This phenomenon is common in applications that consider public health interventions.

As discussed above, the possibility of interference is immaterial when testing the sharp null hypothesis of no effect for any unit. But when the sharp null of no effect is rejected, as in our case, it is natural to ask what can be said about the treatment effect. This magnitude, however, cannot be defined in a straightforward way when a unit's potential outcome depends on other units' treatment assignments. The approach in Rosenbaum (2007) does not assume any particular structure for the type of interference, and is based on test inversion to provide confidence intervals for a particular measure of the benefits of a treatment. More precisely, define a *placebo* or *uniformity* trial as a trial in which units are randomly divided into two groups, but treatment is withheld from all units. In this type of trial, the division of units into groups is merely a labeling of units, since nobody receives any treatment, and therefore the transformed outcomes in the uniformity trial, $\tilde{\mathbf{Y}}^{\mathrm{U}}_{W_0}$, satisfy $\tilde{\mathbf{Y}}^{\mathrm{U}}_{W_0} = \boldsymbol{\alpha}_{W_0}$. Let $T_{\mathrm{U}} := T(\mathbf{D}_{W_0}, \tilde{\mathbf{Y}}^{\mathrm{U}}_{W_0})$ be the value of the statistic in this placebo trial. The key idea is that, although the value of T_{U} is unobservable (because the placebo trial is never performed), in such a trial the null hypothesis of no effect holds by construction and hence the distribution of T_{U} is known.

Define $\Delta = T - T_{\mathrm{U}}$, where the arguments are omitted to ease notation. This magnitude measures the difference in the statistic under the experiment under consideration, T, and the placebo experiment, T_{U}. Hence, if the treatment has no effect, $\Delta = 0$. For example, suppose T is the difference in means. Then $T = \frac{1}{N^+_{W_0}} \sum_{i \in \mathcal{I}_0} \tilde{Y}_i D_i - \frac{1}{N^-_{W_0}} \sum_{i \in \mathcal{I}_0} \tilde{Y}_i (1 - D_i)$ and hence $\Delta = \frac{1}{N^+_{W_0}} \sum_{i \in \mathcal{I}_0} (\tilde{Y}_i - \tilde{Y}^{\mathrm{U}}_i) D_i - \frac{1}{N^-_{W_0}} \sum_{i \in \mathcal{I}_0} (\tilde{Y}_i - \tilde{Y}^{\mathrm{U}}_i)(1 - D_i)$. For this choice of T, Δ is the difference between how

τ, test the null that the (constant) treatment effect is equal to τ and obtain the p-value. Each p-value is plotted in the figure, with darker colors indicating p-values closer to one and whiter colors indicating p-values closer to zero.

Table 6. Ninety-five percent confidence intervals under treatment effect model and interference.

	$W = \hat{W}_0 = [-1.1, 1.1]$	
	No transformation ($p = 0$)	Linear transformation ($p = 1$)
Ages 5–9, HS-targeted causes, post-HS		
RD treatment effect	−2.280	−1.240
Fisher's p-value	0.009	0.006
95% CI, Treatment effect model	[−3.975, −0.575]	[−4.225, −0.800]
95% CI, Interference	[−2.330, −2.218]	[−2.566, −2.449]
N_W^- | N_W^+	43 | 33	43 | 33

Notes: (i) Point estimators are constructed using difference-in-means of untransformed and transformed outcomes, respectively, with a uniform kernel; (ii) randomization p-values and confidence intervals are computed using the methods in Cattaneo et al. (2015) and the new methods presented in this paper (with and without potential outcomes adjustments); (iii) W corresponds to window around zero of the centered running variable $\bar{R}_i = R_i - \bar{r}$; (iv) $N_W^- = \sum_{i=1}^n 1(\bar{r} - w \leq R_i < \bar{r})$, $N_W^+ = \sum_{i=1}^n 1(\bar{r} \leq R_i \leq \bar{r} + w)$; (v) all empirical results are obtained using the implementations described in Cattaneo et al., (2016).

much the treated and control groups deviate (on average) from the zero-effect case. In our case, Δ measures how much bigger the ATE is for treated counties compared to the control counties.[16]

Even though Δ is unobservable, a confidence set for this random variable can be constructed based on observable information. Let κ_1 and κ_2 be some constants satisfying $\kappa_2 < \kappa_1$. Then $\mathbb{P}[T - \kappa_1 \leq \Delta \leq T - \kappa_2] = \mathbb{P}[\kappa_2 \leq T_U \leq \kappa_1]$. Hence, if κ_1 and κ_2 are chosen to be the $\alpha/2$ and $1 - \alpha/2$ quantiles of T_U for some level α, it follows that $\Delta \in [T - \kappa_1, T - \kappa_2]$ with probability $1 - \alpha$. In practice, the values for κ_1 and κ_2 can be recovered from the randomization distribution of T_U, and T is replaced by its observed value.

The main advantage of this approach is that it does not place any restriction on the type of interference that is allowed between units. An alternative approach, described in Bowers, Fredrickson, and Panagopoulos (2013), consists of specifying a parametric model for the potential outcomes, explicitly modeling how the treatment spills from the treated to the control units. This setting allows the researcher to obtain point estimates for treatment effects and to assess how these estimates vary for different degrees of interference.

Table 6 reports Fisherian randomization-based confidence intervals for τ in a constant treatment effect model that assumes SUTVA (described above), and also randomization-based confidence intervals for Δ under arbitrary interference, both reported in our chosen window $\widehat{W}_0 = [-1.1, 1.1]$. The point estimates, already reported in Table 5 above, are also shown for completeness and comparability. The 95 percent confidence interval for τ under SUTVA ranges from roughly −4 to −0.6 under the local constant model, and changes only slightly when estimated under a (local) linear transformation for the outcomes. The confidence intervals that allow for interference are much narrower, do not include zero, and contain the point estimate calculated using difference-in-means—suggesting that the negative results reported in previous sections are "robust" to the presence of arbitrary interference between units. While these confidence intervals are, of course, not strictly

[16] Note that when interference is possible, the average of the differences $\tilde{Y}_i - \tilde{Y}_{Ui}$ for the control group may be nonzero because the treatment may indirectly affect the outcomes of the control units.

Table 7. Rosenbaum sensitivity analysis.

	Window half length (w)						
	$w = 0.3$	$w = 0.5$	$w = 0.7$	$w = 0.9$	$w = 1.1$	$w = 1.3$	$w = 1.5$
Randomization mechanism							
Independent Bernoulli trials	0.0458	0.1028	0.0578	0.0506	0.0098	0.0272	0.0202
Fixed margins	0.0458	0.0954	0.055	0.0456	0.0092	0.0246	0.0188
Upper bound *p*-values							
$\gamma = 0.09; \exp(\gamma) = 1.1$	0.0482	0.109	0.0636	0.0604	0.016	0.031	0.0248
$\gamma = 0.18; \exp(\gamma) = 1.2$	0.0606	0.1296	0.0814	0.0866	0.0288	0.055	0.0444
$\gamma = 0.26; \exp(\gamma) = 1.3$	0.0754	0.1588	0.1126	0.1204	0.0474	0.0904	0.0836
$\gamma = 0.34; \exp(\gamma) = 1.4$	0.0938	0.1968	0.1512	0.1668	0.0774	0.1416	0.129

Notes: (i) Rosenbaum sensitivity analysis conducted as explained in the text, following Rosenbaum (2002b, 2010); Imbens and Rubin (2015); (ii) w corresponds to length of the half windows around zero of the centered running variable $\bar{R}_i = R_i - \bar{r}$, that is, $W = [-w, w]$; (iii) all empirical results are obtained using the implementations described in Cattaneo, Titiunik, and Vazquez-Bare (2016); (iv) the independent Bernoulli trials use $q = \hat{q}$.

comparable, the empirical evidence presented does suggest a statistically significant effect of Head Start on child mortality using randomization-based methods.

Misspecification of the Randomization Mechanism: Rosenbaum Sensitivity Analysis

As a third robustness check, we propose to conduct sensitivity analysis in the local randomization RD framework, following the method in Rosenbaum (2002b). The main idea is to evaluate how inferences about the null hypothesis are affected by the presence of a binary unobservable covariate that changes the probability of receiving treatment. More precisely, assume that the randomization mechanism follows a Bernoulli experiment where the individual probability of treatment $\mathbb{P}(D_i = 1) = q_i = \exp(\gamma U_i)/(1 + \exp(\gamma))$, where $U_i \in \{0, 1\}$ is unobserved. This implies that units with $U_i = 0$ and $U_i = 1$ have different probability of receiving the treatment. The sensitivity analysis considers how different values of γ, which measures the degree of departure from a randomized experiment, affect the randomization p-value. See Rosenbaum (2002b) for further discussion, and Cattaneo, Titiunik, and Vazquez-Bare (2016) for details on implementation.

The results from this sensitivity analysis for different windows can be seen in Table 7. The bounds on the p-values are calculated using a Bernoulli randomization mechanism. The resulting p-values are shown in the first line of the table. For comparison, the second line shows the fixed-margins p-values, which are very close, showing that changing the assignment mechanism does not affect our inferences considerably. The lower panel of Table 7 shows the upper bound for the randomization p-value for different values of γ. We choose values of γ to get $\exp(\gamma)$ in the range {1.1,1.2,1.3,1.4}. These values correspond to an unobserved confounder of increasing "strength": when $\exp(\gamma)$ takes values in {1.1,1.2,1.3,1.4}, we are hypothesizing that the odds of receiving treatment for a treated unit are, respectively, 10, 20, 30, and 40 percent higher than for a control unit.

For our chosen window [−1.1, 1.1], the upper bound is at most 0.077 for the values of γ considered. Hence, even if the unobservable variable U_i could increase the odds of receiving treatment for a treated unit by 40 percent relative to a control unit, the results would still be significant at the ten percent level. The other windows considered in the table seem to be less robust, although most results would remain significant at the 5 percent level even if the odds ratio was increased by

20 percent relative to the case of no hidden bias. These results suggest that the presence of moderately-sized unobservable confounders would not dramatically affect our inferential conclusions.

DISCUSSION

We have discussed and illustrated two alternative approaches to the analysis of RD designs. In the first, most common approach, the conditional regression functions of the potential outcomes given the score are assumed to be continuous around the cutoff and estimation and inference are based on polynomial approximations to these unknown functions and extrapolation at the cutoff.

In the global parametric approach, these regression functions are assumed to have an exact parametric form on the entire support of the running variable. In the flexible parametric approach, a polynomial parametric model is also assumed but this specific model is imposed only in a neighborhood of the cutoff and not on the entire support of the score. Finally, in the nonparametric approach, the shape of the regression functions is left unspecified, and the functions are approximated using nonparametric local polynomial methods, where the neighborhood of approximation is typically chosen optimally to balance bias (which increases as observations far from the cutoff are included in the estimation) and variance (which increases as observations far from the cutoff are discarded to reduce bias), and inference accounts for the resulting misspecification error. Estimation and inference in this framework are always based on large-sample approximations in a super-population framework.

The second framework for RD analysis is one that, instead of relying on continuity of unknown functions and extrapolation, assumes that, in a small window around the cutoff, the treatment is assigned randomly (as it would have occurred in an experiment). Under this assumption, estimation and inference can be based on experimental methods. We emphasized in particular an experimental method where both the potential outcomes and the population of units are seen as fixed and the only randomness comes from the value of the score, which in turn determines treatment assignment. In this Fisherian randomization-based framework, the null distribution of test statistics can be derived exactly from the randomization distribution of the treatment assignment. In the specific RD setting, we do not know exactly what this distribution is, but we can reasonably approximate it by fixed-margins or binomial randomization mechanisms—where, in the latter, the probability of treatment may be estimated from data. Just like choosing a bandwidth is crucial in the large-sample approaches based on continuity, choosing the window where randomization of the treatment is plausible is crucial in the local randomization approach. For this reason, we also introduced and discussed several sensitivity analysis methods in the context of local randomization.

In recent empirical work, it has been common for researchers to adopt a local randomization approach when interpreting results, and a continuity-based approach for estimation and inference using flexible parametric methods. The connection between both approaches, as well as their potential pitfalls, has not been discussed thoroughly, which has often led to misunderstandings. For example, a common misconception is that the bandwidth within which nonparametric local polynomial estimation is conducted is the neighborhood around the cutoff where the treatment can be interpreted to be as-if randomly assigned, or the region where the linear model is correct. But in a continuity-based approach, this is not the interpretation of the bandwidth, as the bandwidth is simply the neighborhood where a polynomial of a given order (usually one or two) is used to approximate the unknown regression function. Indeed, in this framework, the shape of this regression function is most

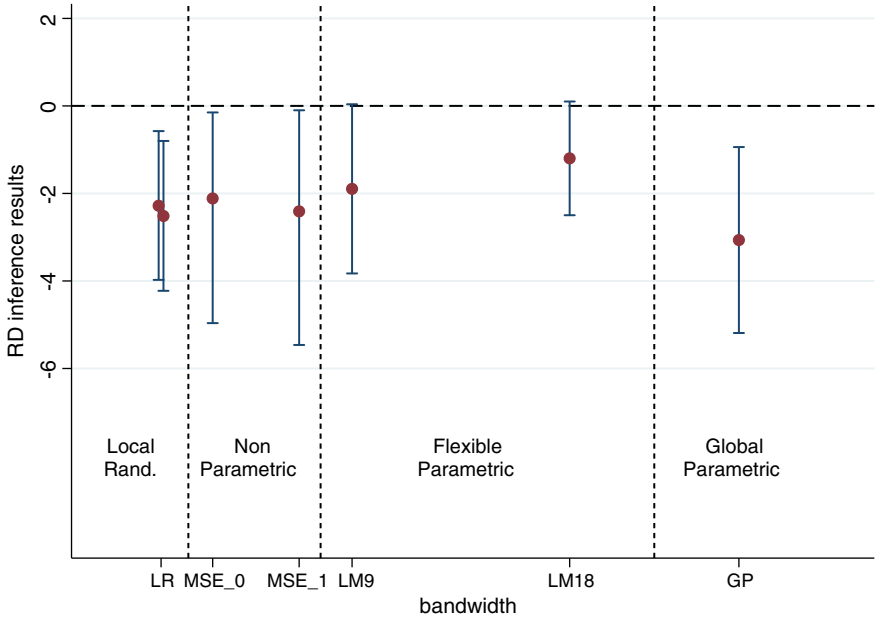
likely not correctly specified, the underlying distribution of treatment assignment is not unrelated to the score or potential outcomes, and the potential outcomes are likely related to the score.

In order for the bandwidth emerging from the continuity-based approach to give a region where a local randomization interpretation is guaranteed, one must adopt a local randomization RD framework and impose more assumptions. In most applications, the continuity-based bandwidth will be too large for such additional assumptions to be plausible, and thus the adoption of a local randomization approach will lead researchers to choose a smaller neighborhood. In the local randomization RD framework, one must assume that there is a window around the cutoff where the treatment is randomly assigned plus an exclusion restriction that prevents the score from affecting the potential outcomes directly—implying that, in the window where randomization is assumed, the potential outcomes are constant functions of the score. Alternatively, as proposed in this paper, one could assume that within this window there is still a relationship between the score and the potential outcomes, but that relationship is separable from the effect of the treatment. For example, if we assume that the potential outcomes are related to the score via a polynomial model whose coefficients are constant among units within each treatment group, then we can transform the potential outcomes to remove the score and adopt Fisherian randomization-inference methods on the transformed outcomes. The empirical implementation of this transformation is comparable to the estimation procedure in continuity-based large-sample methods, where the outcome is regressed on a polynomial of the score within a region set by the choice of bandwidth, and only the difference in intercepts is recorded. Thus, in this sense, our discussion gives a common framework to compare the continuity-based approach with the local randomization approach.

In practice, we see both approaches as complementary. The local randomization approach provides a helpful robustness test on the by now more conventional continuity-based approach, and is particularly helpful when the number of observations near the cutoff is small and the validity of the large-sample approximations employed by the continuity-based framework is more tenuous.

Our application of these methods to the study of the impact of Head Start on child mortality illustrates how this robustness analysis can be implemented. In Figure 4, we summarize the results from both approaches, plotting the point estimate (dots) and 95 percent confidence intervals (bars) as a function of the length of the window or bandwidth used for estimation. The figure is divided in four regions showing, respectively, the RD estimation and inference findings from (i) a local randomization approach based on Fisherian inference with local constant and local linear transformation in the $[-1.1, 1.1]$ window, (ii) a continuity-based nonparametric robust local polynomial approach based on a local constant and a local linear polynomial and MSE-optimal bandwidths ($\hat{h}_{\mathrm{MSE}_0} = 3.235$ for $p = 0$ and $\hat{h}_{\mathrm{MSE}_1} = 6.811$ for $p = 1$), (iii) a continuity-based flexible parametric approach for fixed manually-chosen bandwidths ($\hat{h}_{\mathrm{FP1}} = 9$ and $\hat{h}_{\mathrm{FP2}} = 18$), and (iv) a continuity-based global parametric approach that uses the entire data. As shown in the figure, all methods reach approximately the same conclusion that Head Start decreased the rate of child mortality from HS-targeted causes by about two points, an effect that can be distinguished from zero at the 5 percent level. In particular, the local randomization approach based on Fisherian inference and the large-sample continuity-based approach based on bias-corrected robust nonparametric local polynomial regression, the two recommended methods, yield very consistent results.

Our empirical application thus illustrates a case where the local randomization and continuity-based frameworks yield similar results and conclusions. This may not occur in other applications, so we discuss some of the features of our



Notes: LR corresponds to the point estimate and confidence interval under local randomization using $\hat{h}_{LR} = 1.1$ and a constant specification. CCT0 and CCT1 correspond to the nonparametric specification under constant and linear models, with bandwidths $\hat{h}_{MSE_0} = 3{,}235$ and $\hat{h}_{MSE_1} = 6{,}811$, respectively. LM9 and LM18 correspond to the bandwidths of 9 and 18 used in the flexible parametric linear specification. GP corresponds to a global polynomial or fourth order using the full sample.

Figure 4. Point estimates and confidence intervals.

application that are likely responsible for the agreement between the methods. A crucial assumption of the local randomization framework is that the potential outcomes are either unrelated to the score in the local window—the exclusion restriction originally adopted in Cattaneo et al. (2015)—or are related to it according to a known formula that can be used for adjustment—as in Assumption 5. When this assumption fails but continuity of the potential outcomes holds, the continuity-based framework will lead to valid results but the local randomization framework will not, and thus the results from both methods are likely to disagree.

In the Head Start application, the temporal separation between the running variable and the outcome enhances the plausibility of both the exclusion restriction and the weaker version of this restriction stated in Assumption 5, especially in a sufficiently small window around the cutoff. The running variable is the poverty index calculated using 1960 census results, while the outcome is child mortality between 1973 and 1983. Although a county's poverty level may be strongly related to contemporaneous health outcomes, a county's poverty in 1960 will be more weakly related to health and mortality outcomes occurring 13 to 23 years later—a period long enough for the economic and socio-demographic make-up of the county, and consequently its health outcomes, to change considerably (and possibly non-systematically). This weakened relationship between score and outcome implies that, when looking at a small window around the cutoff where the poverty index in 1960 does not vary by more than a few percentage points, the underlying assumptions required by the local randomization framework are more plausible. Indeed, our empirical analysis (window selection and placebo tests) supports this conclusion.

More generally, the exclusion conditions required by the local randomization framework will be most plausible in applications where there are objective factors weakening the relationship between the score and the outcome near the cutoff. In such cases, the nonparametric continuity-based framework and the local randomization framework are more likely to yield consistent results.

RECOMMENDATIONS FOR PRACTICE

We now offer some general highlights and practical recommendations for the analysis of RD designs based on our recent work and related work available in the literature.

First, global parametric methods based on higher-order polynomials tend to be erratic and perform poorly at boundary points, and hence are unlikely to provide credible and stable results in RD applications. Therefore, we advise against employing this estimation and inference method when analyzing RD designs in applications. Furthermore, standard estimation and inference methods, even local to the cutoff, crucially rely on correct model specification of the regression functions near the RD discontinuity, and as a result are arbitrary because (i) the bandwidth is chosen in an ad hoc manner, and (ii) the impact of misspecification error and bandwidth selection on estimation and inference is (erroneously) disregarded. This implies that the flexible parametric local least-squares regression methods described in Procedure 1 will also underperform in empirical applications, and thus we do not recommend them for applied work.

Second, nonparametric robust bias-corrected local polynomial inference methods account formally for bandwidth selection and misspecification biases. They also permit using the popular MSE-optimal bandwidth, while providing demonstrable improvements in estimation and inference. Finally, they require relatively weak assumptions on the underlying features of the data generating process (e.g., smoothness of the unknown conditional expectations of the potential outcomes), but valid statistical inferences rely on large-sample approximations and extrapolation. All in all, they provide an excellent trade-off between robustness to restrictive assumptions, such as parametric modeling of the unknown conditional expectations, and efficiency. Therefore, we recommend these methods as the default approach for empirical practice. To implement them, the recommended default choices are: (i) local-linear specification ($p = 1$), (ii) triangular kernel ($\mathcal{K}(u) = (1 - |u|)1(|u| \leq 1)$), (iii) second-generation MSE-optimal bandwidth estimator ($h = \hat{h}_{\text{MSE}}$), and (iv) robust bias-corrected inference/confidence intervals. This approach leads to an MSE-optimal point estimator of the RD treatment effect, and valid statistical inference and/or confidence intervals. In addition, inference after robust bias-correction can be made optimal, in a distributional sense, if a different bandwidth is used, though these refinements are more technical and hence beyond the scope of this paper; see Cattaneo and Vazquez-Bare (2016) for more discussion and further references.

Third, local randomization methods will be useful in a more limited set of applications. These methods require the stronger assumption that, possibly after parametric adjustment of the outcome variables, a local randomization assumption holds. This assumption is stronger than the usual assumption of continuity/smoothness of conditional expectations, and hence it may not hold in all applications, but it can be empirically falsified as we discussed and illustrated in this paper. The key advantage of this assumption is that, whenever it holds, it allows for finite-sample exact inference methods and other useful methods from the analysis of experiments literature. Therefore, whenever this assumption is believed plausible in empirical applications, an array of additional estimation and inference methods becomes available, which

complements the more standard local polynomial methods and gives further robustness checks and complementary empirical evidence.

In particular, we recommend employing local randomization methods in two types of scenarios: (i) when the exclusion restriction or Assumption 5 is plausible (in addition to the other required assumptions), and (ii) when the running variable is discrete—that is, when multiple observations have the same score value (see also Dong, 2015; Lee & Card, 2008). As mentioned above, scenario (i) is more likely to apply in cases where there are objective factors that weaken the relationship between the score and the outcome. In this scenario, our recommendation is to use local randomization methods not as primary analysis but rather as a secondary analysis to establish the robustness of the results obtained using local polynomial methods within the continuity-based framework. In contrast, the continuity-based methods are not directly applicable when the running variable is discrete without further assumptions. In this case, we recommend local randomization methods as the primary analysis focusing only on the observations closest to the cutoff. When the score is discrete, the local randomization framework has the advantage that the choice of window can be avoided, as the smallest possible window is the window that includes the cutoff point (where all observations are assigned to treatment) and the score value immediately below it (where all observations are assigned to control). Moreover, the methods in this Fisherian framework will yield exact inferences even if the sample size in this window is very small, which may occur in some applications. This approach changes the estimand of interest, but in cases where the running variable is inherently discrete, this is expected and reasonable.

To summarize, if the running variable is continuous, we recommend using continuity-based local polynomial methods for analysis, and local randomization methods as a robustness check whenever applicable. If the running variable is discrete, we recommend using local randomization methods as the primary analysis (and possibly continuity-based methods under additional assumptions).

CONCLUSION

We offered a comprehensive discussion of the main inference methods currently available in the literature for the analysis of RD designs, and applied them to a substantive case study. Motivated by the influential work of Ludwig and Miller (2007), who employed global and flexible parametric methods in a continuity-based RD framework, we reexamined the effect of Head Start on child mortality employing two main modern inference approaches: nonparametric robust local polynomial inference within a continuity-based RD framework, and finite-sample exact inference within a local randomization RD framework. We also introduced and discussed methodological extensions to the latter framework allowing for parametric transformation of outcomes and sensitivity analyses. Applying both frameworks to the re-analysis of the effect of Head Start on child mortality, we showed that this effect is strongly consistent across the different methods considered.

An extension of our methods to fuzzy RD designs is given in the Supporting Information Appendix.[17] The methodological results discussed in this paper can also be used in other related RD settings, including multi-running variable RD designs (e.g., Papay, Willett, & Murnane, 2011), geographic RD designs (e.g., Keele

[17] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

& Titiunik, 2015), and multi-cutoff RD designs (e.g., Bertanha, 2017; Cattaneo et al., 2016). We do not discuss these extensions here to conserve space.

MATIAS D. CATTANEO is an Associate Professor in the Department of Economics and in the Department of Statistics at the University of Michigan, 238 Lorch Hall, 611 Tappan Avenue, Ann Arbor, MI 48109-1220 (e-mail: cattaneo@umich.edu).

ROCÍO TITIUNIK is the James Orin Murfin Associate Professor in the Department of Political Science at the University of Michigan, 5700 Haven Hall, 505 South State Street, Ann Arbor, MI 48109-1045 (e-mail: titiunik@umich.edu).

GONZALO VAZQUEZ-BARE is a Ph.D. Candidate in the Department of Economics at the University of Michigan, 238 Lorch Hall, 611 Tappan Avenue, Ann Arbor, MI 48109-1220 (e-mail: gvazquez@umich.edu).

ACKNOWLEDGMENTS

We thank Martha Bailey, Jake Bowers, Sebastian Calonico, Max Farrell, Xinwei Ma, Doug Miller, Ariel Pihl, Tomás Rau, and Elizabeth Stuart for comments and suggestions. We also thank Jens Ludwig and Doug Miller for sharing their original dataset. Finally, we thank the Editor in Chief, the Methods Section Editor, and four anonymous reviewers for their insightful and detailed comments, which greatly improved our manuscript. Financial support from the National Science Foundation (SES 1357561) is gratefully acknowledged.

REFERENCES

Administration for Children and Families, U.S. Department of Health and Human Services. (2010, January). Head Start impact study. Final report. Washington, DC.

Bertanha, M. (2017). Regression discontinuity design with many thresholds. Working paper, University of Notre Dame, IN, USA.

Bowers, J., Fredrickson, M. M., & Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. Political Analysis, 21, 97–124.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2016). On the effect of bias estimation on coverage accuracy in nonparametric inference. Journal of the American Statistical Association, forthcoming.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2017). Coverage error optimal confidence intervals for regression discontinuity designs. Working paper, University of Michigan, MI, USA.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2016). Regression discontinuity designs using covariates. Working paper, University of Michigan, MI, USA.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression discontinuity designs. Stata Journal. Forthcoming.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014a). Robust data-driven inference in the regression-discontinuity design. Stata Journal, 14, 909–946.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014b). Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82, 2295–2326.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015a). Optimal data-driven regression discontinuity plots. Journal of the American Statistical Association, 110, 1753–1769.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015b). rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. R Journal, 7, 38–51.

Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. Econometrica, 83, 2453–2483.

Cattaneo, M. D., & Escanciano, J. C. (2017). Regression discontinuity designs: Theory and applications. Advances in econometrics (Vol. 38). Emerald Group Publishing, forthcoming.

Cattaneo, M. D., Frandsen, B., & 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, 1–24.

Cattaneo, M. D., Jansson, M., & Ma, X. (2016a). Simple local regression distribution estimators with an application to manipulation testing. Working paper, University of Michigan, MI, USA.

Cattaneo, M. D., Jansson, M., & Ma, X. (2016b). rddensity: Manipulation testing based on density discontinuity. Working paper, University of Michigan, MI, USA.

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

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2016). Inference in regression discontinuity designs under local randomization. Stata Journal, 16, 331–367.

Cattaneo, M. D., & Vazquez-Bare, G. (2016). The choice of neighborhood in regression discontinuity designs. Observational Studies, 2, 134–146.

Cerulli, G., Dong, Y., Lewbel, A., & Poulsen, A. (2017). Testing stability of regression discontinuity models. In M. D. Cattaneo & J. C. Escanciano (Eds.), Regression discontinuity designs: Theory and applications. Advances in Econometrics (Vol. 38). Bingley, UK: Emerald Group Publishing.

Cheng, M.-Y., Fan, J., & Marron, J. S. (1997). On automatic boundary corrections. Annals of Statistics, 25, 1691–1708.

Chiang, H. D., & Sasaki, Y. (2016). Causal inference by quantile regression kink designs. Working paper, MD, USA.

Congress of the United States, Congressional Budget Office, Microeconomic Studies Division. (2013). Growth in means-tested programs and tax credits for low-income households. Washington, DC.

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, 636–654.

Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. Journal of Applied Econometrics, 30, 422–446.

Dong, Y., & Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. Review of Economics and Statistics, 97, 1081–1092.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. Statistical Science, 19, 676–685.

Fisher, R. A. (1935). Design of experiments. New York, NY: Hafner.

Frandsen, B. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In M. D. Cattaneo & J. C. Escanciano (Eds.), Regression discontinuity designs: Theory and applications. Advances in econometrics (Vol. 38). Bingley, UK: Emerald Group Publishing.

Ganong, P., & Jager, S. (2016). A permutation test for the regression kink design. Working Paper, Harvard University, MA, USA.

Gelman, A., & Imbens, G. W. (2014). Why high-order polynomials should not be used in regression discontinuity designs. NBER Working Paper 20405. New York, NY: National Bureau of Economic Research.

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

Ho, D. E., & Imai, K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. Journal of the American Statistical Association, 101, 888–900.

Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945–960.

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. Review of Economic Studies, 79, 933–959.

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

Imbens, G. W., & Rosenbaum, P. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. Journal of the Royal Statistical Society, Series A, 168, 109–126.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge, UK: Cambridge University Press.

Jales, H., & Yu, Z. (2017). Identification and estimation using a density discontinuity approach. In M. D. Cattaneo & J. C. Escanciano (Eds.), Regression discontinuity designs: Theory and applications. Advances in econometrics (Vol. 38). Bingley, UK: Emerald Group Publishing.

Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. Political Analysis, 23, 127–155.

Keele, L. J., Titiunik, R., & Zubizarreta, J. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. Journal of the Royal Statistical Society: Series A, 178, 223–239.

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

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. Journal of Econometrics, 142, 655–674.

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

Lehmann, E. L. (1998). Nonparametrics: Statistical methods based on ranks. New York, NY: Prentice Hall.

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, 159–208.

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

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics, 161, 203–207.

Pihl, A. M. (2016). Head Starts impacts on mothers: Regression discontinuity evidence from the program's early years. Working paper, UC-Davis, CA, USA.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. Statistical Science, 17, 286–327.

Rosenbaum, P. R. (2002b). Observational studies. New York, NY: Springer.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. Journal of the American Statistical Association, 102, 191–200.

Rosenbaum, P. R. (2010). Design of observational studies. New York, NY: Springer.

Sekhon, J., & Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In M. D. Cattaneo & J. C. Escanciano (Eds.), Regression discontinuity designs: Theory and applications. Advances in econometrics (Vol. 38). Bingley, UK: Emerald Group Publishing.

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

Wand, M., & Jones, M. (1995). Kernel smoothing. Boca Raton, FL: Chapman & Hall/CRC.

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

APPENDIX

This supplemental appendix reports additional empirical results not included in the main paper to conserve space. It also discusses the extension of the new randomization-based methods to the fuzzy RD design introduced in the article.

LOCAL RANDOMIZATION: CONCEPTUAL ISSUES

In this section, we discuss formal definitions and ideas behind the concept of local randomization in the RD context. In particular, we discuss the interpretation of RD designs as local experiments building on the arguments in Cattaneo, Frandsen, and Titiunik (2016) and Sekhon and Titiunik (2017), and making explicit connections to the local randomization interpretation presented by Lee (2008).

In a very influential piece, Lee (2008) advocated for interpreting RD designs as local experiments near the cutoff. This view, later expanded in Lee and Lemieux (2010), argues that in RD designs where units lack perfect control over the score value they receive, the variation in treatment assignment induced by the RD assignment rule can be interpreted to be as good as randomized. Lee (2008) proposed a behavioral model in which the assumption of "lack of perfect manipulation of the running variable" translates into continuity of the units' unobservable characteristics (or "types") at the cutoff.

The analogy between RD designs and local experiments has had the beneficial effect of encouraging researchers to test for equality or "balance" of the distribution of predetermined covariates between treated and control units at the cutoff, analogously to the way in which covariate balance is tested in experiments. Such falsification tests are now common in RD empirical analysis, and have contributed to the credibility of many RD applications. However, the analogy between RD and experiments has also created some confusion, in particular about when and where continuity conditions on potential outcomes and densities of unobservable types lead to actual experimental conditions for any given finite sample.

The framework in Lee (2008) justifies using the analogy between RD and experiments only heuristically. The reason is that the conditions in the Lee framework are all based on continuity of relevant functions. These conditions guarantee the validity of the continuity-based nonparametric approaches described in our main paper, but do not justify the use of techniques from randomized experiments. Importantly, as discussed by Sekhon and Titiunik (2016), the continuity conditions in Lee (2008) do not follow by design. By itself, the RD treatment assignment rule $D_i = \mathbb{1}(R_i \geq \bar{r})$ imposes no restrictions on the shape or properties of the potential outcomes regression functions or the density of unobservable characteristics. These restrictions must be imposed in addition to the RD treatment rule, which is why the credibility of the RD design ranks below the credibility of actual experiments, where the key independence condition holds by design.

A contribution of our paper is to give conditions under which RD designs can be interpreted and analyzed as experiments, not heuristically but rather in a precise sense. A crucial obstacle to offer a precise analogy between RD designs and experiments is the intermediate role of the running variable, which is absent in an experiment. In most RD applications, the running variable is an important determinant of the outcomes of interest. For example, in our Head Start application, a county's poverty is likely strongly related to other characteristics of the county that affect the mortality outcomes of interest, such as income, demographic composition, etc. Moreover, by affecting families' ability to access health care services, the poverty index could have a direct effect on child mortality. This illustrates the general fact that the RD running variable may both correlate with predetermined characteristics that are related to the outcome of interest, and have a "direct" effect on the potential

outcomes. The interpretation of RD designs as local experiments in a precise (rather than heuristic) sense hinges on the assumption that such relationships between the score and potential outcomes do not exist—the exclusion restriction in Cattaneo et al. (2015)—or can be removed by adjustment—our Assumption 5 in the main paper.

Note that in an actual experiment, there is also an intermediate variable akin to the score in a RD design, but this variable is by definition unrelated to the potential outcomes. In every experiment, a random chance device is used to assign units to treatment and control. Often, this device is a pseudo-random number generator in a computer that assigns a random number to every unit in the study. This number is then used as the basis for the treatment assignment. For example, one could assign a treatment with probability 50 percent by assigning a uniform random number between 0 and 100 to all units, and then assigning the treatment to those units whose number is above 50. It is clear, however, that the intermediate random number in an actual experiment is by construction entirely unrelated to the potential outcomes or any other characteristics of the units. Indeed, in most experiments, the experimental units do not know what their random number is, they only know the final treatment/control assignment. See Sekhon and Titiunik (2017) for further discussion.

In other words, in an actual experiment, the exclusion restriction between the score and the potential outcomes holds by construction. In contrast, in RD designs, this exclusion restriction is not guaranteed by the RD assignment mechanism. (In fact, in our Head Start example, the poverty index is chosen as the running variable precisely because of its predominant role in determining the overall health-related and socio-economic outcomes of municipalities, making the exclusion restriction extremely implausible.) Thus, if researchers wish to analyze RD designs using experimental methods, they must make this assumption explicitly (or an assumption such as Assumption 5 that allows for adjustment). Moreover, as discussed at length by Sekhon and Titiunik (2017), even the assumption that the value of the RD score is randomly assigned among units near the cutoff is insufficient to guarantee that this exclusion restriction holds. The reason is that although such random assignment prevents the score from being related to predetermined characteristics of the units, it does not preclude the score from affecting the potential outcomes via post-treatment channels.

To clarify these concepts further, we provide some formalization. Local randomization may be analyzed in two alternative frameworks. The first approach assumes that the potential outcomes are drawn from a super-population, and are therefore random. This is the framework used in the main paper for discussing continuity-based methods. In this context, randomization means that the potential outcomes are statistically independent of treatment assignment. Local randomization occurs when this independence only holds inside some window.

Definition 1 (Local independence of treatment: super-population)**.**

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid R_i \in W_0, \qquad W_0 = [\bar{r} - w, \bar{r} + w]$$

Sekhon and Titiunik (2017) point out that this local independence condition does not imply that the potential outcomes are unrelated to the running variable. (An equivalent way to define local randomization is to use distribution functions, so Definition 1 can be recast as $\mathbb{P}[D_i = 1 | Y_i(1), Y_i(0), R_i \in W_0] = \mathbb{P}[D_i = 1 | R_i \in W_0]$.) In other words, the local independence assumption in Definition 1 is not enough to guarantee that the exclusion restriction discussed above holds. Indeed, there are cases where both potential outcomes are functions of R and local independence as stated in Definition 1 holds (see Sekhon & Titiunik, 2017, for an example).

A stronger version of the assumption states local independence between potential outcomes and the score variable, instead of between potential outcomes and the treatment assignment indicator. This stronger assumption can be stated as

Definition 2 (Local independence of score: super-population)**.**

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp R_i \mid R_i \in W_0$$

or equivalently

$$\mathbb{P}[R_i \leq r | Y_i(1), Y_i(0), R_i \in W_0] = \mathbb{P}[R_i \leq r | R_i \in W_0], \qquad r \in \mathscr{R}. \tag{A.1}$$

Unlike the local independence condition in Definition 1, the stronger local independence in Definition 2 implies that $\mathbb{E}[Y_i(d)|R_i, R_i \in W_0] = \mathbb{E}[Y_i(d)|R_i \in W_0]$, $d = 0, 1$, so the exclusion restriction does hold and the conditional expectations are flat (as functions of R_i) inside the window. Importantly, note that neither Definition 1 nor Definition 2 are satisfied in the Lee (2008) framework, as continuity of conditional expectations or distributions of potential outcomes (or covariates) does not imply either type of local independence.

In sum, when adopting the super-population local randomization framework as in Definition 1, the exclusion restriction must be explicitly adopted in addition to this assumption. Alternatively, researchers can adopt Definition 2, which essentially assumes that the score variable plays no intermediate role, and hence implies that the exclusion restriction holds.

The second framework in which local randomization can be analyzed is the Fisherian framework that we employ in the main paper. In this context, potential outcomes are seen as non-random, and the only randomness comes from the treatment assignment, or in this case, the running variable (which deterministically defines the treatment in sharp RD designs). In the main paper, we define local randomization in the following way:

Definition 3 (Local randomization: finite sample)**.** *There exists a window* $W_0 = [\bar{r} - w, \bar{r} + w]$, $w > 0$, *such that the following holds:*

1. *Non-Random Potential Outcomes.* $\mathbf{y}(\mathbf{d}, \mathbf{r})$ *are fixed.*
2. *Unconfoundedness.* $\mathbb{P}(\mathbf{R}_{W_0} \leq \mathbf{r}\,;\, \mathbf{y}(\mathbf{d}, \mathbf{r})) = \mathbb{P}(\mathbf{R}_{W_0} \leq \mathbf{r})$*, for all vectors* $\mathbf{r} \in \mathscr{R}_{W_0}$.
3. *Mechanism.* $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$ *is known for all vectors* $\mathbf{d} \in \mathscr{D}_{W_0}$.

Part 2 of this definition imposes a condition equivalent to equation A.1 (Definition 2), but seeing the potential outcomes as non-random. In addition, the finite-sample version requires knowledge about the probability of receiving treatment, which is required to calculate all possible values of the treatment assignment vector. These requirements, together with an assumption on the functional relationship between the potential outcomes and the running variable (see Assumption 5 in the main paper and also next section) yield a scenario that mimics a randomized experiment. In other words, the Fisherian finite-sample local randomization framework proposed by Cattaneo et al. (2015) and extended in the main paper contains an explicit exclusion restriction. As mentioned before, the behavioral model proposed by Lee (2008) does not justify this interpretation.

RANDOMIZATION *P*-VALUES: NUMERICAL IMPLEMENTATION

As explained in the main text, the finite-sample exact p-value can be found theoretically by calculating the probability that the test statistic exceeds the value that it takes in the observed sample. Because the potential outcomes are seen as fixed, the

only source of randomness is the treatment assignment, and hence this probability can be calculated by enumerating all the treatment assignment vectors that are possible under the known randomization mechanism used to randomly assign the treatment in the observed sample. For a numerical example, we refer the reader to Imbens and Rubin (2015).

The formula for the finite-sample exact p-value is:

$$\mathbb{P}\left(T\left(\mathbf{D}_{W_0}, \boldsymbol{\alpha}^0_{W_0}\right) \geq T_{obs}\right) = \sum_{\mathbf{d} \in \mathscr{D}_{W_0}} \mathbb{1}\left(T\left(\mathbf{d}, \boldsymbol{\alpha}^0_{W_0}\right) \geq T_{obs}\right) \cdot \mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d}),$$

where $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$ is the known distribution stated in Definition 3 (and Assumption 4 in the main paper).

Even for moderate sample sizes, however, the total number of possible values that the treatment vector $\mathbf{D}$ can take will be prohibitively large. As a simple illustration, the number of possible values of $\mathbf{D}$ for a sample of size 20 with 10 treated units and 10 control units is 184,756, whereas for a sample of 30 with 15 treated units and 15 control units it becomes 155,117,520. Hence, in practice, the finite sample distribution of the statistic is approximated by drawing a random sample from the known distribution $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$. In the particular case of fixed-margins randomization, this random sample is obtained by drawing random permutations of the treatment vector in the observed sample—that is, by sampling the observed treatment assignment vector without replacement.

In the particular case that $\mathbb{P}(\mathbf{D}_{W_0} = \mathbf{d})$ follows a fixed-margins randomization, the randomization-based p-value is obtained in practice with the following procedure:

1. Choose a test statistic T.
2. Using the observed sample, calculate the observed value of the test statistic, T_{obs}.
3. Obtain a permutation of the treatment assignment by reshuffling the ones and zeros in the vector $\mathbf{D}$. Call this permuted vector $\mathbf{D}^\pi_1$.
4. Calculate the value of the test statistic, T^π_1, for this permuted treatment assignment $\mathbf{D}^\pi_1$.
5. Repeat steps 3 and 4 a large number of times S, for example $S = 1000$, to obtain a vector of length S with the values of the test statistic for all the permutations of the treatment assignment. Collect all the values of the statistic under each permutation in a vector $\mathbf{T}^\pi$.
6. Obtain the finite sample p-value by calculating the number of times an element in $\mathbf{T}^\pi$ exceeds the observed statistic T_{obs}, and dividing that number by S.

ROSENBAUM'S METHOD FOR CONFIDENCE INTERVALS UNDER INTERFERENCE

The idea in Rosenbaum (2015b) is to look at the magnitude $\Delta := T - T_{\mathrm{U}}$, where T is some test statistic and T_{U} is the value of this statistic that would be observed when the treatment is withheld from all units (i.e., in a uniformity trial). This magnitude asks whether there is a greater tendency for treated subjects to have higher responses than controls relative to what would have been observed in a uniformity trial. The quantity T_{U} is unobserved because in practice the uniformity trial is never performed. However, because in a uniformity trial the null hypothesis holds by construction, the distribution of T_{U} (which is determined by the treatment assignment) is known. For some statistics such as the Wilcoxon-Mann-Whitney rank sum statistic, this distribution can be obtained without reference to the data. More generally, the distribution can be simulated in the same way in which finite-simple

p-values are obtained, that is, by permuting the treatment and calculating the value of the statistic in each step (as described in the previous section). Note that:

$$\begin{aligned}\mathbb{P}(\Delta > T - c_\alpha) &= \mathbb{P}(T - T_{\mathrm{U}} > T - c_\alpha) \\ &= \mathbb{P}(-T_{\mathrm{U}} > -c_\alpha) \\ &= \mathbb{P}(T_{\mathrm{U}} < c_\alpha) \\ &= 1 - \mathbb{P}(T_{\mathrm{U}} \geq c_\alpha)\end{aligned}$$

so if c_α is chosen to ensure that $\mathbb{P}(T_{\mathrm{U}} \geq c_\alpha) = \alpha$ (i.e., c_α is the critical value from the distribution of T_{U}, which is known), we obtain that $\Delta > T - c_\alpha$ with probability $1 - \alpha$, so the set $(T - c_\alpha; \infty)$ can be seen as a $1 - \alpha$ level confidence interval for Δ. This confidence set can be constructed using the known distribution of T_{U} and the observed value of the statistic T. In practice, the critical value c_α is obtained by setting a value of α, simulating the distribution of the statistic under the null, and finding the number c_α such that the proportion of cases in which the statistic falls above c_α is equal to α. The value of T in the confidence interval simply comes from the observed value of the statistic in the sample. As an illustration, if $T = 10$ and $\mathbb{P}(T_{\mathrm{U}} \geq 4) = 0.05$ (so that $\alpha = 0.05$ and $c_{0.05} = 4$) we have that $\mathbb{P}(\Delta > 6) = 0.95$ so we can assert that $\Delta > 6$ with 95 percent confidence. This reasoning extends straightforwardly to the two-sided case described in the paper.

ADDITIONAL EMPIRICAL RESULTS

In this section, we present additional empirical results for the Head Start application. First, we give descriptive statistics for the full sample as well as for different subsamples of counties with score near the RD cutoff. Second, we perform several formal falsification tests to provide empirical evidence supporting the validity of the RD design. Finally, we offer further empirical results using nonparametric local polynomial and local randomization methods.

Basic Descriptive Statistics

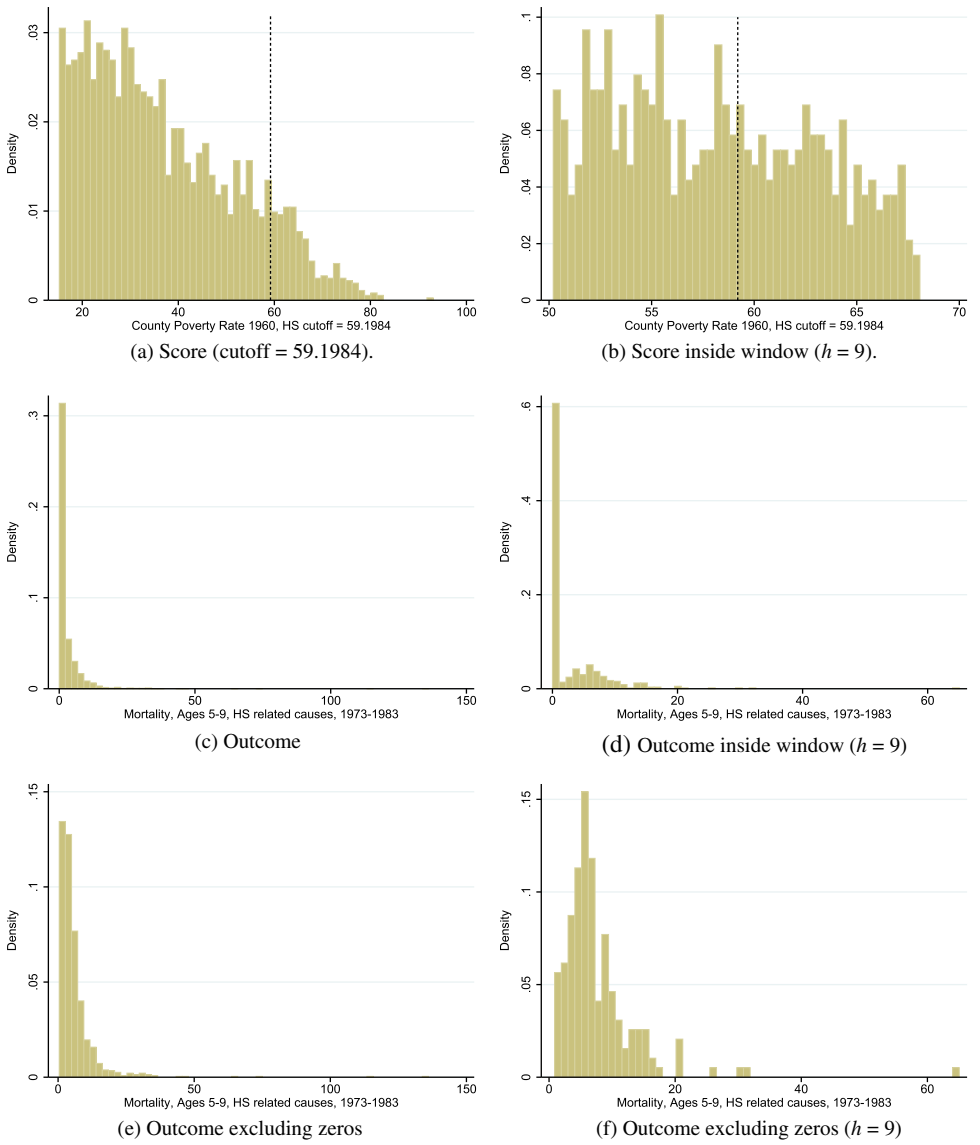
Figures A1 and A2 show histograms of the running variable and the outcome of interest for the full sample as well as for different windows around the cutoff. Visually, the running variable does not seem to have a discontinuity around the cutoff; this is confirmed by the falsification tests conducted below. The plot for the outcome reveals a big mass point at zero, both outside and inside the window. Even conditional on being positive, the outcome variable concentrates in low values.

Descriptive statistics for these two variables can be seen in Table A1. The poverty index ranges roughly from 15 to 93 with a mean of 33.6. We can see that the cutoff is located close to the 90th percentile of the score distribution, as expected since the program implementation (RD design) is based on a fixed number of the poorest U.S. counties. On the other hand, the outcome variable (child mortality) has a mean of 2.25 but a median equal to zero, which is consistent with the strong asymmetry seen in Figure A3.

Finally, Tables A2 through A5 report difference-in-means of the outcome variable and pre-intervention covariates for the full sample and different windows around the cutoff.

Falsification Tests: Additional Results

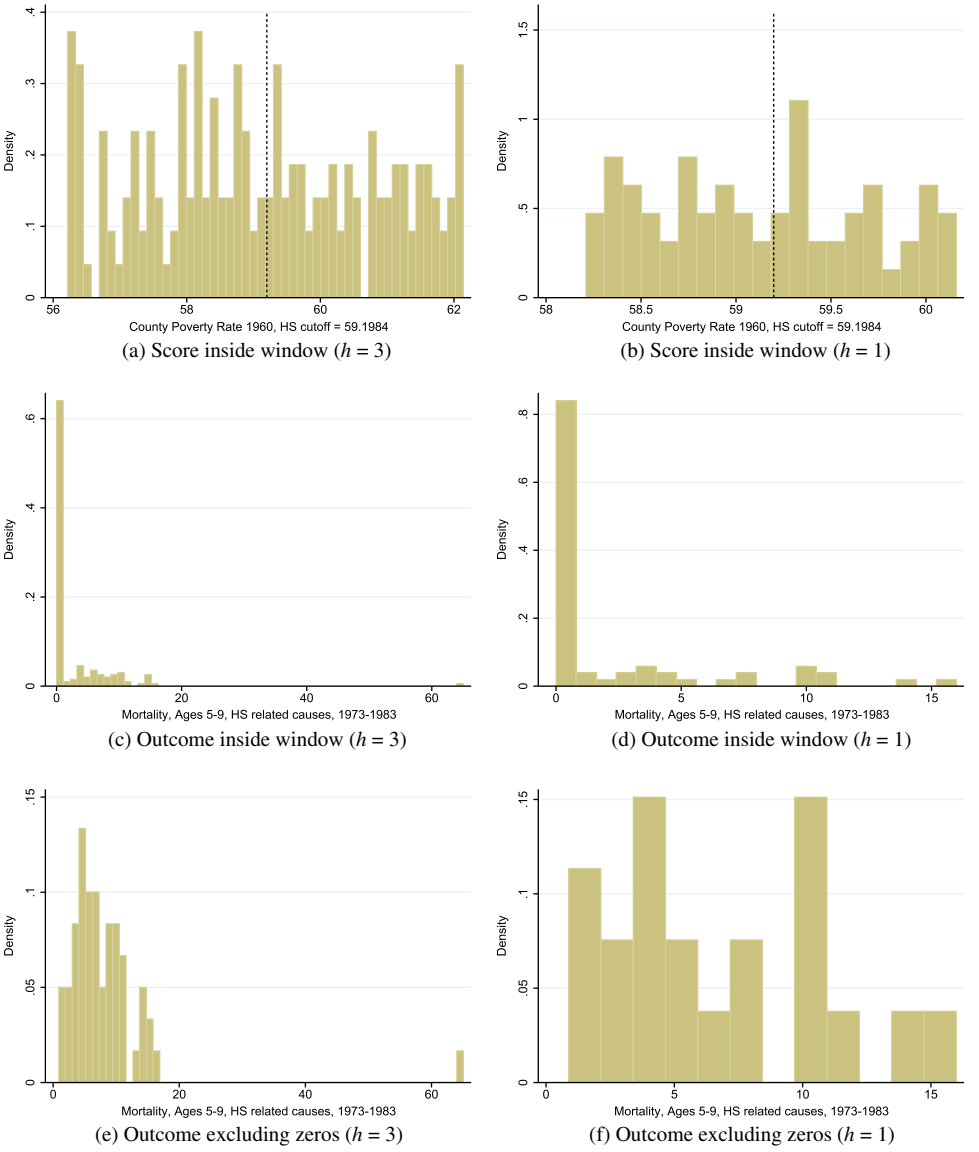
To check for continuity away from the cutoff on the outcome variable, Figure A4 presents two RD plots constructed using the Integrated Mean Square Error (IMSE)



Notes: the left column shows the histogram of the score (poverty index), the outcome (mortality, ages five to nine, HS-related causes) and the outcome conditional on being positive for the full sample. The right column replicates the same histograms but inside a window around the cutoff with bandwidth $h = 9$. The dashed line indicates the cutoff (59.1984).

Figure A1. Running variable and outcome variable.

optimal number of disjoint bins to approximate the underlying regression functions, under repeated sampling. Panel (a) employs evenly spaced bins over the support of the running variable, while panel (b) employs quantile spaced bins. The idea is to compare a global polynomial fit (smooth approximation) to a local sample-means fit over disjoint bins (discontinuous approximation), with the goal of identifying possible discontinuities away from the cutoff. These optimal RD plots are fully data-driven and follow the recent results in Calonico et al. (2015). The figure shows



(a) Score inside window ($h = 3$)

(b) Score inside window ($h = 1$)

(c) Outcome inside window ($h = 3$)

(d) Outcome inside window ($h = 1$)

(e) Outcome excluding zeros ($h = 3$)

(f) Outcome excluding zeros ($h = 1$)

Notes: the left column shows the histogram of the score (poverty index), the outcome (mortality, ages five to nine, HS-related causes) and the outcome conditional on being positive inside a window with bandwidth $h = 3$. The right column replicates the same histograms but inside a window with bandwidth $h = 1$. The dashed line indicates the cutoff (59.1984).

Figure A2. Running variable and outcome variable (cont.)

a fairly consistent picture for the control units ($R_i < \bar{r} = 59.1984$), where both the global smooth polynomial fit and the local discontinuous sample means exhibit a very similar behavior. For the treatment group ($R_i \geq \bar{r}$), on the other hand, the two approaches show some noticeable differences that may require further analysis. We do not find any graphical or formal evidence of additional discontinuities over the support of the running variable.

Table A1. Descriptive statistics for running variable and outcome.

	Poverty index	Mortality, ages 5–9, HS-targeted
Mean	36.787	2.254
Standard deviation	15.350	5.726
Min	15.209	0.000
10th percentile	18.737	0.000
25th percentile	24.139	0.000
50th percentile	33.615	0.000
75th percentile	47.426	2.828
90th percentile	59.765	6.658
Max	93.072	136.054
Obs	2,804.000	2,783.000

Nonparametric Local Polynomial Methods

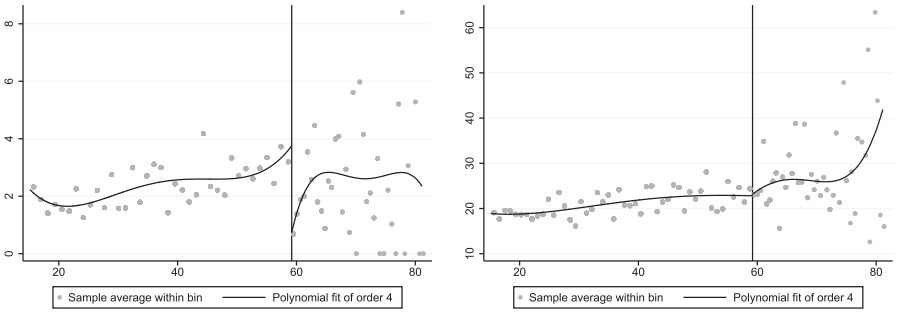
A natural robustness check when implementing RD methods is to report estimation and inference results for different bandwidth choices. Table A6 shows the results for four different bandwidth selection methods: $\hat{h}_{\texttt{CER}}$ corresponds to the coverage-error-optimal bandwidth from Calonico, Cattaneo and Farrell (2017), $\hat{h}_{\texttt{MSE}}$ corresponds to the MSE-optimal bandwidth used in the paper and suggested by Calonico, Cattaneo, and Titiunik (2014), and $\hat{h}_{\texttt{FP1}}$ and $\hat{h}_{\texttt{FP2}}$ correspond to the bandwidths used in the flexible parametric approach. All the results are provided for three polynomial models: constant, linear, and quartic.

In all models except for the constant one (in which they are equal by construction), the CER-optimal bandwidth is smaller than the MSE-optimal bandwidth. Nevertheless, the main empirical results do not change dramatically for the different bandwidths for the constant and linear models, revealing negative and statistically significant results. The point estimates roughly range from around -2.4 to -1.0, all of them significantly different from zero at 5-percent level. In addition, Tables A7 and A8 report details on the placebo tests on the pre-intervention covariates, which also show very stable and robust results, most of them revealing large p-values (with a few exceptions that could be false positives due to multiple testing issues). The quartic model, on the other hand, shows generally larger and unstable results. This model, however, is not recommended in practice, and previous work has shown that high-order polynomials do not perform well in these contexts (see e.g., Gelman & Imbens, 2014).

In sum, our findings using nonparametric local polynomial methods do not seem to be sensitive to the choice of the bandwidth, and appear to be quite robust.

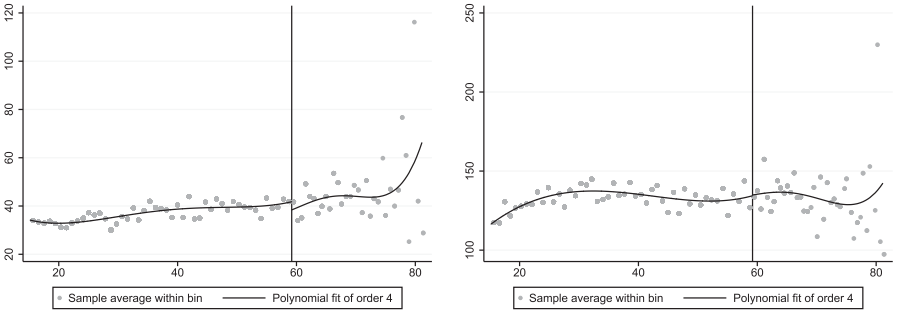
Local Randomization Methods

In this section, we explore how the results obtained under the local randomization approach are affected by different choices of the window length. For example, different data-driven selected lengths can be obtained by using different statistics for the balance tests. Figure A5 shows the minimum p-value for covariate balance tests using the Kolmogorov-Smirnov, t-test, rank sum, and Hotelling statistics. The resulting chosen windows in these cases have half-length equal to $w = 1.1$, $w = 1.3$, $w = 1.5$, and $w = 2.7$, respectively. The smallest window corresponds to the Kolmogorov-Smirnov test, which is intuitive because the Kolmogorov-Smirnov is more demanding as it requires the whole distribution to be balanced between



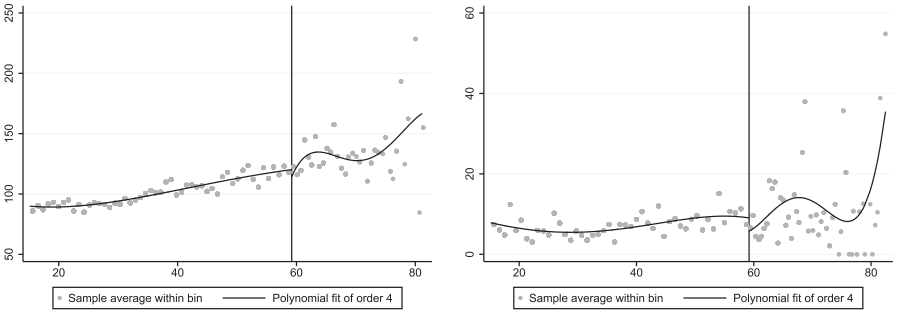
(a) Ages 5–9, HS-related causes, post-HS

(b) Ages 5–9, injuries, post-HS



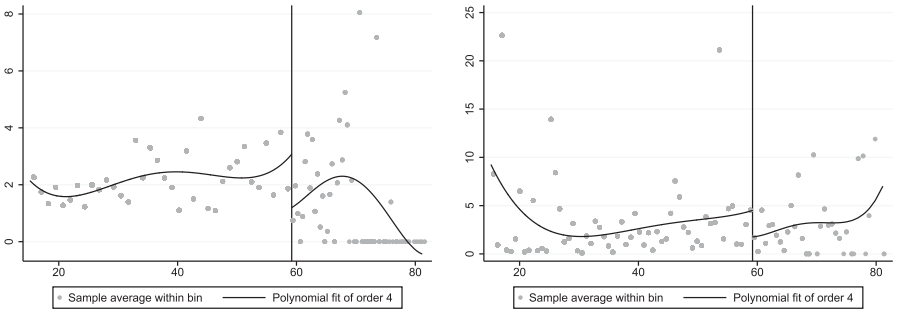
(c) Ages 5–9, all causes, post-HS

(d) Ages 25+, HS-related causes, post-HS



(e) Ages 5–9, injuries, post-HS

(f) Ages 5–9, HS-related causes, pre-HS



(g) Whites, Ages 5–9, HS-related causes, post-HS

(h) Blacks, Ages 5–9, HS-related causes, post-HS

Notes: Data-driven RD plots using Mimicking Variance Optimal Number of Bins. See Calonico, Cattaneo, and Titiunik (2015) for details.

Figure A3. RD Plots Main and Placebo Outcomes.

Table A2. Summary statistics and difference-in-means, full sample.

	Control group ($R_i < 59.1984$)			Treatment group ($R_i \geq 59.1984$)			Difference-in-means		
	Obs.	Sample mean	Std. err.	Obs.	Sample mean	Std. err.	Diff-in-means	Std. err.	p-Value
Main Variables									
Ages 5–9, HS-targeted causes (Y_i)	2489	2.234	0.117	294	2.422	0.263	0.188	0.288	0.515
1960 Poverty Index (R_i)	2504	33.280	0.241	300	66.054	0.322	32.774	0.402	0.000
Falsification Variables									
Ages 5–9, injuries, post-HS	2489	20.618	0.336	294	25.960	1.146	5.342	1.195	0.000
Ages 5–9, all causes, post-HS	2489	36.333	0.431	294	42.759	1.404	6.427	1.469	0.000
Ages 25+, HS-targeted, post-HS	2489	132.387	0.732	294	134.648	1.778	2.261	1.923	0.240
Ages 25+, Injuries, post-HS	2489	98.651	0.527	294	131.890	2.385	33.239	2.442	0.000
Ages 5–9, HS-targeted, pre-HS	2504	6.840	0.350	300	10.675	0.843	3.835	0.913	0.000
Whites, Ages 5–9, HS-targeted, post-HS	2489	2.087	0.129	294	1.706	0.313	−0.381	0.339	0.262
Blacks, Ages 5–9, HS-targeted, post-HS	2103	3.443	0.733	267	2.647	0.375	−0.796	0.824	0.334
1960 Census Variables									
County population	2504	41.527	2.476	300	17.567	0.946	−23.961	2.651	0.000
% Attending school, Ages 14–17	2486	84.433	0.338	294	80.946	0.633	−3.487	0.718	0.000
% Attending school, Ages 5–34	2487	0.549	0.001	294	0.569	0.003	0.020	0.003	0.000
% High-school or more, Ages 25+	2486	34.804	0.193	294	19.407	0.269	−15.397	0.331	0.000
Population, Ages 14–17	2487	2.692	0.140	294	1.531	0.079	−1.161	0.161	0.000
Population, Ages 5–34	2487	19.419	1.111	294	8.812	0.503	−10.607	1.219	0.000
Population, Ages 25+	2487	23.289	1.486	294	8.411	0.422	−14.877	1.545	0.000
% Urban population	2492	30.972	0.541	294	13.390	1.055	−17.581	1.185	0.000
% Black population	2492	7.886	0.258	294	33.911	1.538	26.025	1.560	0.000

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A3. Summary statistics and difference-in-means, $|R_i - \bar{r}| \leq 9$, $\bar{r} = 59.1984$.

	Control group ($R_i < 59.1984$)			Treatment group ($R_i \geq 59.1984$)			Difference-in-means		
	Obs.	Sample mean	Std. err.	Obs.	Sample mean	Std. err.	Diff-in-means	Std. err.	p-Value
Main Variables									
Ages 5–9, HS-targeted causes (Y_i)	309	2.989	0.333	215	2.298	0.309	−0.691	0.454	0.129
1960 Poverty Index (R_i)	310	54.602	0.146	217	63.142	0.165	8.540	0.220	0.000
Falsification Variables									
Ages 5–9, injuries, post-HS	309	22.759	0.928	215	25.794	1.400	3.036	1.680	0.072
Ages 5–9, all causes, post-HS	309	39.887	1.222	215	42.101	1.671	2.214	2.070	0.285
Ages 25+, HS-targeted, post-HS	309	131.939	1.720	215	136.249	2.100	4.310	2.714	0.113
Ages 25+, Injuries, post-HS	309	117.478	1.362	215	130.489	2.942	13.010	3.242	0.000
Ages 5–9, HS-targeted, pre-HS	310	9.501	0.996	217	10.013	0.929	0.512	1.362	0.707
Whites, Ages 5–9, HS-targeted, post-HS	309	2.546	0.343	215	1.693	0.355	−0.853	0.494	0.085
Blacks, Ages 5–9, HS-targeted, post-HS	287	4.231	1.573	200	2.473	0.433	−1.758	1.632	0.282
1960 Census Variables									
County population	310	18.689	0.925	217	18.851	1.230	0.162	1.539	0.916
% Attending school, Ages 14–17	309	82.072	0.594	215	81.408	0.706	−0.665	0.923	0.472
% Attending school, Ages 5–34	309	0.552	0.003	215	0.564	0.003	0.012	0.004	0.003
% High-school or more, Ages 25+	309	22.596	0.272	215	20.179	0.294	−2.417	0.400	0.000
Population, Ages 14–17	309	1.533	0.075	215	1.608	0.102	0.075	0.127	0.554
Population, Ages 5–34	309	9.173	0.498	215	9.328	0.653	0.155	0.821	0.850
Population, Ages 25+	309	9.561	0.432	215	9.095	0.544	−0.467	0.695	0.503
% Urban population	309	18.814	1.112	215	14.348	1.230	−4.466	1.658	0.007
% Black population	309	20.888	1.051	215	31.822	1.610	10.934	1.923	0.000

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A4. Summary statistics and difference-in-means, $|R_i - \bar{r}| \leq 3$, $\bar{r} = 59.1984$.

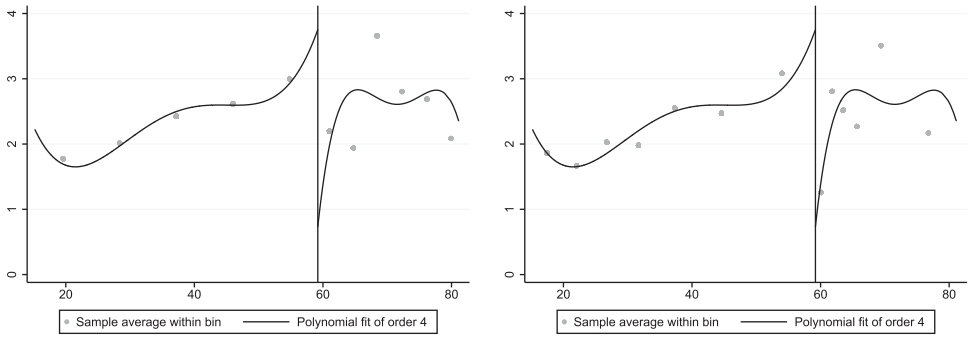
	Control group ($R_i < 59.1984$)			Treatment group ($R_i \geq 59.1984$)			Difference-in-means		
	Obs.	Sample mean	Std. err.	Obs.	Sample mean	Std. err.	Diff-in-means	Std. err.	p-Value
Main Variables									
Ages 5–9, HS-targeted causes (Y_i)	96	3.198	0.776	84	1.917	0.414	−1.281	0.880	0.148
1960 Poverty Index (R_i)	96	57.750	0.092	85	60.676	0.099	2.926	0.135	0.000
Falsification Variables									
Ages 5–9, injuries, post-HS	96	23.054	1.567	84	24.494	2.570	1.440	3.010	0.633
Ages 5–9, all causes, post-HS	96	41.568	2.298	84	40.388	2.911	−1.180	3.709	0.751
Ages 25+, HS-targeted, post-HS	96	133.632	3.097	84	136.925	4.008	3.293	5.065	0.517
Ages 25+, Injuries, post-HS	96	120.108	2.088	84	126.939	4.679	6.831	5.124	0.185
Ages 5–9, HS-targeted, pre-HS	96	9.876	1.158	85	6.071	0.869	−3.805	1.447	0.009
Whites, Ages 5–9, HS-targeted, post-HS	96	2.709	0.796	84	1.721	0.590	−0.988	0.991	0.320
Blacks, Ages 5–9, HS-targeted, post-HS	88	2.645	0.719	77	2.067	0.685	−0.578	0.993	0.561
1960 Census Variables									
County population	96	18.111	1.900	85	20.398	1.772	2.287	2.598	0.380
% Attending school, Ages 14–17	96	80.845	1.409	84	81.390	1.544	0.546	2.090	0.794
% Attending school, Ages 5–34	96	0.555	0.006	84	0.570	0.004	0.015	0.007	0.037
% High-school or more, Ages 25+	96	21.231	0.475	84	20.838	0.483	−0.393	0.678	0.562
Population, Ages 14–17	96	1.482	0.148	84	1.729	0.150	0.247	0.211	0.244
Population, Ages 5–34	96	8.748	1.020	84	10.041	0.930	1.293	1.380	0.350
Population, Ages 25+	96	8.900	0.841	84	9.958	0.808	1.058	1.166	0.365
% Urban population	96	15.934	1.825	84	15.556	1.928	−0.378	2.655	0.887
% Black population	96	23.618	1.973	84	28.230	2.416	4.612	3.119	0.141

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A5. Summary statistics and difference-in-means, $|R_i - \bar{r}| \leq 1, \bar{r} = 59.1984$.

	Control group ($R_i < 59.1984$)			Treatment group ($R_i \geq 59.1984$)			Difference-in-means		
	Obs.	Sample mean	Std. err.	Obs.	Sample mean	Std. err.	Diff-in-means	Std. err.	p-Value
Main Variables									
Ages 5–9, HS-targeted causes (Y_i)	35	3.220	0.801	29	0.909	0.419	−2.311	0.904	0.014
1960 Poverty Index (R_i)	35	58.678	0.048	30	59.651	0.054	0.973	0.072	0.000
Falsification Variables									
Ages 5–9, injuries, post-HS	35	23.865	2.421	29	24.094	3.050	0.229	3.894	0.953
Ages 5–9, all causes, post-HS	35	41.745	3.567	29	39.787	3.772	−1.959	5.191	0.707
Ages 25+, HS-targeted, post-HS	35	127.545	3.757	29	139.373	6.028	11.828	7.103	0.102
Ages 25+, Injuries, post-HS	35	119.537	4.104	29	119.751	3.948	0.214	5.694	0.970
Ages 5–9, HS-targeted, pre-HS	35	7.269	1.359	30	8.017	1.643	0.748	2.132	0.727
Whites, Ages 5–9, HS-targeted, post-HS	35	1.547	0.632	29	1.217	0.871	−0.330	1.076	0.760
Blacks, Ages 5–9, HS-targeted, post-HS	35	5.110	1.619	28	1.329	1.208	−3.781	2.020	0.066
1960 Census Variables									
County population	35	23.376	4.422	30	23.704	3.240	0.327	5.482	0.953
% Attending school, Ages 14–17	35	83.003	1.299	29	82.814	0.994	−0.189	1.636	0.908
% Attending school, Ages 5–34	35	0.566	0.009	29	0.563	0.007	−0.003	0.011	0.804
% High-school or more, Ages 25+	35	21.300	0.826	29	21.786	0.666	0.486	1.061	0.648
Population, Ages 14–17	35	1.875	0.329	29	1.996	0.268	0.121	0.424	0.776
Population, Ages 5–34	35	11.484	2.402	29	11.829	1.749	0.346	2.971	0.908
Population, Ages 25+	35	11.144	1.931	29	11.889	1.468	0.745	2.426	0.760
% Urban population	35	17.349	3.279	29	20.403	3.953	3.055	5.136	0.554
% Black population	35	26.740	3.295	29	25.214	3.794	−1.526	5.025	0.762

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management



(a) Evenly Spaced Bins
$J_- = 5 \; ; \; J_+ = 6$

(b) Quantile Spaced Bins
$J_- = 7 \; ; \; J_+ = 6$

Notes: (i) solid blue lines depict 4th order polynomial fits using control and treated units separately, and (iv) dots depict sample average of outcome variable within each bin.

Figure A4. RD Plots using IMSE-optimal Approximation, Head Start Data.

groups. The window chosen in the main paper is therefore the most conservative one.

Table A9 shows the inference results for different window lengths, including the four chosen using the four statistics mentioned above and $w = 0.9$, which is an even more conservative choice than the ones obtained via balance tests. For the model without adjustment, the point estimates range between around -1.6 and -3.1, all of them statistically significant at the 5-percent level. Additionally, the placebo tests yield mostly insignificant results, with only few exceptions. Hence, the results seem quite robust to different window lengths.

For the linear adjustment, the results are also reasonably robust but less stable in terms of point estimation magnitudes. Specifically, the point estimates vary from around -1 to -4, with increasing p-values for larger windows. Some placebo tests reveal low p-values for the larger windows, where the local randomization assumption is less plausible. To better understand the effects of linear transformations on the outcome variable, Figure A6 shows the scatter plot of the outcome of interest (child mortality post-HS) against the re-centered running variable, together with the linear adjustment model for $w = 1.1$ (panel a) and $w = 1.3$ (panel b) transformations. The graph clearly shows that the change in the results when increasing the window length from 1.1 to 1.3 is mostly driven by a single observation with a very high value (i.e., an outlier). Given the low number of observations used near the cutoff, the fitted slope is very sensitive to outliers and this might make the estimates and p-values change abruptly, a phenomenon that affects the constant model less severely. For this reason, the linear transformation in the local randomization framework should be used with caution, and should be accompanied by a thorough graphical analysis and robustness checks (specially when the number of observations in the chosen window is very small).

For completeness, in Table A10 we also report analogous results using large-sample (instead of randomization-based) inference methods.

Summary of Empirical Results

Finally, for completeness, the information used to construct Figure 4 in the main paper is given in Table A11.

Table A6. Robust nonparametric local polynomial methods.

	Constant model ($p = 0$)				Linear model ($p = 1$)				Quartic model ($p = 4$)			
	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$
Ages 5–9, HS-targeted causes, post-HS												
RD treatment effect	−2.114	−2.114	−1.059	−0.662	−3.273	−2.409	−2.182	−1.567	−1.755	−3.068	−3.302	−3.384
Robust p-value	0.037	0.037	0.048	0.047	0.011	0.042	0.027	0.035	0.527	0.197	0.547	0.005
$N_W^- \mid N_W^+$	98 | 92	98 | 92	309 | 215	671 | 283	150 | 132	234 | 180	309 | 215	671 | 283	191 | 161	296 | 213	309 | 215	671 | 283
h	3.235	3.235	9.000	18.000	4.581	6.810	9.000	18.000	5.730	8.651	9.000	18.000
Falsification Tests, robust p-values												
Ages 5–9, injuries, post-HS	0.880	0.880	0.960	0.547	0.925	0.728	0.787	0.964	0.759	0.918	0.767	0.694
Ages 5–9, all causes, post-HS	0.400	0.400	0.388	0.700	0.387	0.516	0.504	0.416	0.546	0.800	0.432	0.766
Ages 25+, HS-targeted causes, post-HS	0.806	0.806	0.716	0.352	0.753	0.866	0.649	0.770	0.301	0.330	0.255	0.375
Ages 25+, Injuries, post-HS	0.705	0.705	0.764	0.216	0.685	0.731	0.645	0.771	0.781	0.907	0.934	0.840
Ages 5–9, HS-targeted causes, pre-HS	0.242	0.242	0.044	0.135	0.508	0.468	0.378	0.013	0.152	0.075	0.162	0.536
Whites, Ages 5–9, HS-targeted causes, post-HS	0.297	0.297	0.282	0.158	0.224	0.311	0.256	0.231	0.624	0.695	0.640	0.171
Blacks, Ages 5–9, HS-targeted causes, post-HS	0.153	0.153	0.500	0.165	0.312	0.742	0.585	0.328	0.667	0.173	0.839	0.205

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) "robust p-values" are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014); (iii) $\hat{h}_{\texttt{MSE}}$ corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico, Cattaneo, and Titiunik (2016); Calonico et al. (2014); (iv) $N_W^- = \sum_{i=1}^n \mathbb{1}(h - \bar{r} \leq R_i \leq \bar{r})$, $N_W^+ = \sum_{i=1}^n \mathbb{1}(\bar{r} \leq R_i \leq \bar{r} + h)$; (v) all results are obtained using the software implementations described in Calonico, Cattaneo, Farrell, and Titiunik, (2017, and references therein).

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A7. Robust nonparametric local polynomial methods, main and placebo outcomes.

	Constant model ($p = 0$)				Linear model ($p = 1$)			
	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$
Ages 5–9, HS-targeted causes, post-HS								
RD treatment effect	−2.114	−2.114	−1.059	−0.662	−3.273	−2.409	−2.182	−1.567
Robust p-value	0.037	0.037	0.048	0.047	0.011	0.042	0.027	0.035
$N_W^- \mid N_W^+$	98 | 92	98 | 92	309 | 215	671 | 283	150 | 132	234 | 180	309 | 215	671 | 283
h	3.235	3.235	9.000	18.000	4.581	6.810	9.000	18.000
Ages 5–9, injuries, post-HS								
RD treatment effect	1.439	1.439	2.198	2.598	0.248	1.133	0.169	1.586
Robust p-value	0.880	0.880	0.960	0.547	0.925	0.728	0.787	0.964
$N_W^- \mid N_W^+$	124 | 111	124 | 111	309 | 215	671 | 283	139 | 122	211 | 169	309 | 215	671 | 283
h	3.932	3.932	9.000	18.000	4.211	6.261	9.000	18.000
Ages 5–9, all causes, post-HS								
RD treatment effect	−2.164	−2.164	0.508	1.700	−5.016	−3.501	−3.615	−1.227
Robust p-value	0.400	0.400	0.388	0.700	0.387	0.516	0.504	0.416
$N_W^- \mid N_W^+$	108 | 103	108 | 103	309 | 215	671 | 283	139 | 122	215 | 170	309 | 215	671 | 283
h	3.546	3.546	9.000	18.000	4.269	6.346	9.000	18.000
Ages 25+, HS-targeted causes, post-HS								
RD treatment effect	2.228	2.228	3.643	3.597	2.710	2.032	2.090	4.018
Robust p-value	0.806	0.806	0.716	0.352	0.753	0.866	0.649	0.770
$N_W^- \mid N_W^+$	131 | 114	131 | 114	309 | 215	671 | 283	181 | 154	281 | 203	309 | 215	671 | 283
h	4.060	4.060	9.000	18.000	5.418	8.055	9.000	18.000

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) "robust p-values" are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014); (iii) $\hat{h}_{\texttt{MSE}}$ corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014, 2016); (iv) $N_W^- = \sum_{i=1}^n \mathbb{1}(h - \bar{r} \leq R_i \leq \bar{r})$, $N_W^+ = \sum_{i=1}^n \mathbb{1}(\bar{r} \leq R_i \leq \bar{r} + h)$; (v) all results are obtained using the software implementations described in Calonico, Cattaneo, Farrell, and Titiunik (2017, and references therein).

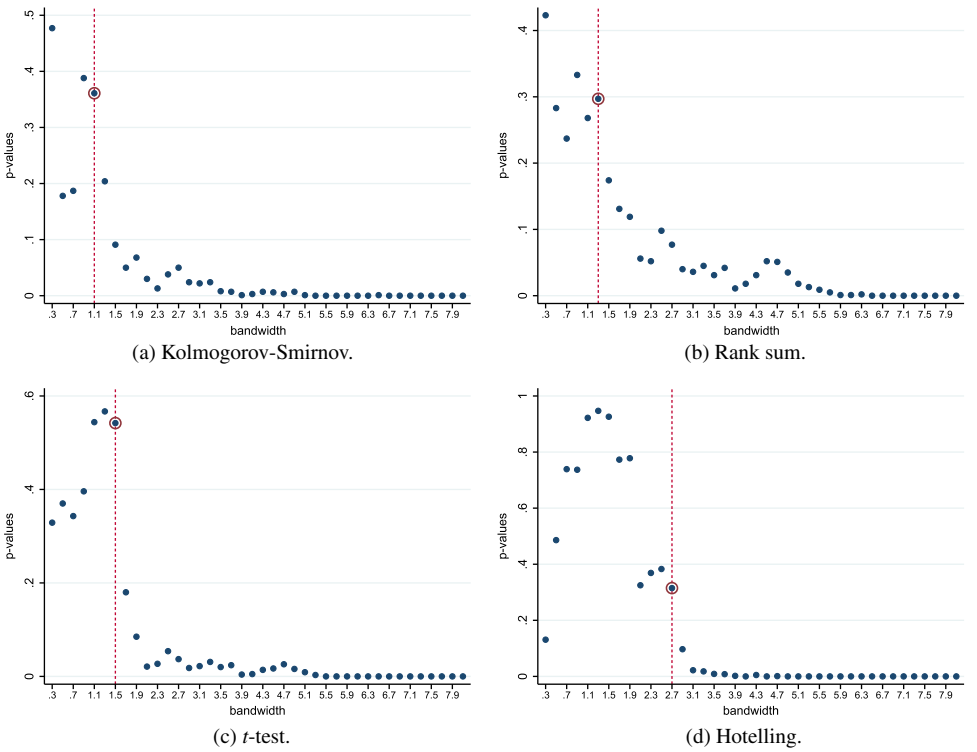
Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A8. Robust nonparametric local polynomial methods, main and placebo outcomes (cont'd).

	Constant model ($p = 0$)				Linear model ($p = 1$)			
	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$	$h = \hat{h}_{\texttt{CER}}$	$h = \hat{h}_{\texttt{MSE}}$	$h = \hat{h}_{\texttt{FP1}}$	$h = \hat{h}_{\texttt{FP2}}$
Ages 25+, Injuries, post-HS								
RD treatment effect	1.345	1.345	10.814	14.734	−1.411	0.052	1.551	5.301
Robust p-value	0.705	0.705	0.764	0.216	0.685	0.731	0.645	0.771
$N_W^- \mid N_W^+$	73 | 64	73 | 64	309 | 215	671 | 283	138 | 120	207 | 167	309 | 215	671 | 283
h	2.295	2.295	9.000	18.000	4.177	6.209	9.000	18.000
Ages 5–9, HS-targeted causes, pre-HS								
RD treatment effect	−2.061	−2.061	−0.628	1.052	1.493	−2.533	−3.540	−2.486
Robust p-value	0.242	0.242	0.044	0.135	0.508	0.468	0.378	0.013
$N_W^- \mid N_W^+$	80 | 73	80 | 73	310 | 217	674 | 287	100 | 96	164 | 138	310 | 217	674 | 287
h	2.644	2.644	9.000	18.000	3.284	4.884	9.000	18.000
Whites, Ages 5–9, HS-targeted causes, post-HS								
RD treatment effect	−1.016	−1.016	−0.928	−0.708	−1.674	−1.317	−1.269	−1.210
Robust p-value	0.297	0.297	0.282	0.158	0.224	0.311	0.256	0.231
$N_W^- \mid N_W^+$	164 | 138	164 | 138	309 | 215	671 | 283	173 | 140	268 | 193	309 | 215	671 | 283
h	4.905	4.905	9.000	18.000	5.092	7.571	9.000	18.000
Blacks, Ages 5–9, HS-targeted causes, post-HS								
RD treatment effect	−1.901	−1.901	−1.901	−1.516	−2.572	−1.080	−1.145	−2.257
Robust p-value	0.153	0.153	0.500	0.165	0.312	0.742	0.585	0.328
$N_W^- \mid N_W^+$	285 | 200	285 | 200	287 | 200	596 | 259	130 | 113	201 | 157	287 | 200	596 | 259
h	8.976	8.976	9.000	18.000	4.310	6.356	9.000	18.000

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) "robust p-values" are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014); (iii) $\hat{h}_{\texttt{MSE}}$ corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014, 2016); (iv) $N_W^- = \sum_{i=1}^n \mathbb{1}(h - \bar{r} \leq R_i \leq \bar{r})$, $N_W^+ = \sum_{i=1}^n \mathbb{1}(\bar{r} \leq R_i \leq \bar{r} + h)$; (v) all results are obtained using the software implementations described in Calonico, Cattaneo, Farrell, and Titiunik (2017, and references therein).

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management



(a) Kolmogorov-Smirnov. (b) Rank sum.

(c) *t*-test. (d) Hotelling.

Notes: minimum *p*-value as a function of window length for Kolmogorov-Smirnov, t-test, rank sum and Hotelling statistics.

Figure A5. Window selection.

EXTENSION TO FUZZY RD DESIGNS

In a fuzzy RD design, treatment assignment is no longer a deterministic function of the score because of imperfect compliance. Let $\mathbf{d}(\mathbf{r})$ denote the vector of potential treatment status as a function of $\mathbf{r}$. The observed treatment status is $\mathbf{D} = \mathbf{d}(\mathbf{R})$. Let $Z_i = \mathbb{1}(R_i \geq 0)$ be the treatment assignment for unit i and collect these variables in a vector $\mathbf{Z}$ taking values $\mathbf{z} \in \mathscr{D}$, which will be used as an instrument for $\mathbf{D}$.

In principle, the potential outcomes are a function $\mathbf{y}(\mathbf{d}, \mathbf{r}, \mathbf{z})$. A key assumption in experiments with imperfect compliance, however, is that assignment to treatment does not have a direct effect on potential outcomes—only taking or not taking the treatment should have an effect, not the assignment itself. In our context, this means that being above the cutoff can only affect potential outcomes through $\mathbf{r}$ and by changing the treatment status $\mathbf{d}$. This is usually known as the exclusion restriction in instrumental variables models (see, e.g., Imbens & Rubin, 2015). Note that this exclusion restriction is different from the exclusion restriction mentioned above: the restriction in instrumental variables model refers to the exclusion of the treatment assignment indicator Z from the potential outcomes, while the previous exclusion restriction referred to the exclusion of the score variable R. We state this restriction as follows:

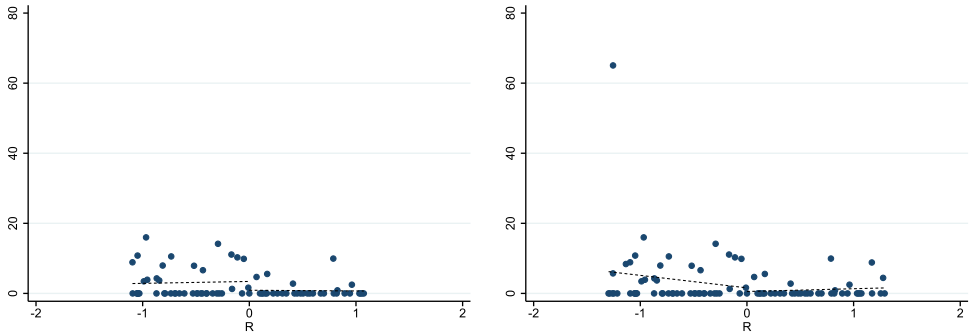
Assumption 1 (Finite Population and Assignment Mechanism)**.** *There exists a window* $W_0 = [\underline{r}, \bar{r}]$ *with* $\underline{r} < 0 < \bar{r}$ *such that the following holds:*

Table A9. Local randomization methods: finite-sample inference.

	No adjustment ($p = 0$)					Linear adjustment ($p = 1$)				
	$h = 0.9$	$h = 1.1$	$h = 1.3$	$h = 1.5$	$h = 2.7$	$h = 0.9$	$h = 1.1$	$h = 1.3$	$h = 1.5$	$h = 2.7$
Ages 5–9, HS-targeted causes, post-HS										
RD treatment effect	−1.908	−2.280	−3.105	−3.089	−1.632	−3.631	−2.515	−1.041	−1.147	−3.999
Fisher's p-value	0.039	0.008	0.031	0.020	0.081	0.000	0.003	0.668	0.553	0.000
$N_W^- \mid N_W^+$	32 | 27	43 | 33	51 | 38	53 | 40	81 | 74	32 | 27	43 | 33	51 | 38	53 | 40	81 | 74
h	0.900	1.100	1.300	1.500	2.700	0.900	1.100	1.300	1.500	2.700
Falsification Tests, Fisher's *p*-values										
Ages 5–9, injuries, post-HS	0.954	0.698	0.855	0.963	0.585	0.354	0.192	0.896	0.560	0.983
Ages 5–9, all causes, post-HS	0.796	0.348	0.190	0.109	0.652	0.364	0.147	0.103	0.067	0.244
Ages 25+, HS-targeted causes, post-HS	0.059	0.285	0.973	0.976	0.412	0.876	0.032	0.001	0.001	0.816
Ages 25+, Injuries, post-HS	0.989	0.956	0.870	0.825	0.203	0.393	0.567	0.388	0.379	0.408
Ages 5–9, HS-targeted causes, pre-HS	0.823	0.936	0.619	0.597	0.020	0.159	0.247	0.071	0.108	0.560
Whites, Ages 5–9, HS-targeted causes, post-HS	0.955	0.473	0.335	0.230	0.425	0.022	0.604	0.488	0.311	0.008
Blacks, Ages 5–9, HS-targeted causes, post-HS	0.127	0.070	0.119	0.144	0.300	0.436	0.322	0.062	0.025	0.000

Notes: (i) Point estimator is constructed using difference in means of unadjusted and adjusted outcomes, respectively, with a uniform kernel; (ii) Fisher's p-values are obtained by taking 10,000 draws from the distribution of the treatment assignment assuming a fixed-margins randomization; (iii) $\hat{h}_{\mathrm{LR}}$ corresponds to the bandwidth obtained by the method described in Cattaneo et al. (2015). Results were obtained using the implementations described in Cattaneo, Titiunik, and Vazquez-Bare (2016).

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management



(a) Linear adjustment forw= 1.1 (b) Linear adjustment forw= 1.3.

Notes: panel (a) shows the linear adjustment for the outcome of interest with $w = 1.1$; panel (b) shows the same graph but for $w = 1.3$. The figure reveals that the difference in point estimates and inference between the two cases is largely driven by a single outlier.

Figure A6. Sensitivity of linear adjustment model.

1. $\mathbf{y}(\mathbf{d}, \mathbf{r}, \mathbf{z})$ *are fixed.*
2. $\mathbb{P}(\mathbf{R}_{W_0} \leq \mathbf{t}; \mathbf{Y}(\mathbf{d}, \mathbf{r}, \mathbf{z})) = \mathbb{P}(\mathbf{R}_{W_0} \leq \mathbf{t})$ *for* $\mathbf{t} \in \mathscr{R}_{W_0}$.
3. $\mathbb{P}(\mathbf{Z}_{W_0} = \mathbf{z})$ *is known for all* $\mathbf{z} \in \mathscr{D}_{W_0}$.
4. *For any pair* $\underline{t}$, $\bar{t}$ *such that* $\underline{r} < \underline{t} < 0 < \bar{t} < \bar{r}$, $\mathbb{P}(D_i = 1 \mid R_i = \bar{t}) > \mathbb{P}(D_i = 1 \mid R_i = \underline{t})$ *for all* i *with* $R_i \in W_0$.

Assumption 1 requires no selection and that the researcher know the probability distribution of the treatment assignment. The only difference is the addition of condition (*4*), which states that the probability of treatment below the cutoff is strictly lower than the probability of being treated above the cutoff. Note that in a sharp design this condition holds automatically, since the probability of treatment is an indicator function and jumps from zero to one at the cutoff.

Assumption 2 (Exclusion Restriction on the Treatment Assignment)**.** $\mathbf{y}(\mathbf{d}, \mathbf{r}, \mathbf{z}) = \mathbf{y}(\mathbf{d}, \mathbf{r})$ *for all* $\mathbf{d}$.

Assumption 2 allows us to drop the argument $\mathbf{z}$ from the potential outcomes, which become $\mathbf{y}(\mathbf{d}, \mathbf{r})$ as before.

Assumption 3 (Transformed Outcomes)**.** *For all* i *such that* $R_i \in W_0$*, the following holds:*

1. *There exists a function* $\phi : \mathscr{Y} \times \mathscr{R} \times \mathscr{D} \to \mathbb{R}$ *such that the adjusted potential outcomes only depend on* $\mathbf{r}$ *through* $\mathbf{z}$*, that is,*

$$\phi(y_i(\mathbf{d}, \mathbf{r}), \mathbf{d}, \mathbf{r}) = \tilde{y}_i(\mathbf{d}, \mathbf{z}) \quad \forall\, \mathbf{r} \in \mathscr{R}_{W_0},\ \mathbf{d} \in \mathscr{D}_{W_0}$$

2. *The potential treatment status only depends on* $\mathbf{z}$*, that is,*

$$\mathbf{d}(\mathbf{r}) = \mathbf{d}(\mathbf{z}) \quad \forall\, \mathbf{r} \in \mathscr{R}_{W_0}$$

Finally, we assume that SUTVA holds for both the adjusted potential outcomes and treatment status.

Assumption 4 (Local SUTVA)**.** *For all* i *with* $R_i \in W_0$*,*

1. $d_i(\mathbf{z}) = d_i(z_i)$ *for all* $\mathbf{z} \in \mathscr{D}_{W_0}$
2. $\tilde{y}_i(\mathbf{d}) = \tilde{y}_i(d_i)$ *for all* $\mathbf{d} \in \mathscr{D}_{W_0}$.

Table A10. Local randomization methods: large-sample inference.

	No adjustment ($p = 0$)					Linear adjustment ($p = 1$)				
	$h = 0.9$	$h = 1.1$	$h = 1.3$	$h = 1.5$	$h = 2.7$	$h = 0.9$	$h = 1.1$	$h = 1.3$	$h = 1.5$	$h = 2.7$
Ages 5–9, HS-targeted causes, post-HS										
RD treatment effect	−1.908	−2.280	−3.105	−3.089	−1.632	−3.631	−2.515	−1.041	−1.147	−3.999
Asymptotic p-value	0.033	0.005	0.029	0.025	0.100	0.078	0.160	0.698	0.637	0.003
$N_W^- \mid N_W^+$	32 | 27	43 | 33	51 | 38	53 | 40	81 | 74	32 | 27	43 | 33	51 | 38	53 | 40	81 | 74
h	0.900	1.100	1.300	1.500	2.700	0.900	1.100	1.300	1.500	2.700
Falsification Tests, Fisher's p-values										
Ages 5–9, injuries, post-HS	0.962	0.696	0.858	0.965	0.582	0.629	0.464	0.944	0.747	0.986
Ages 5–9, all causes, post-HS	0.797	0.340	0.187	0.117	0.618	0.640	0.430	0.383	0.315	0.435
Ages 25+, HS-targeted causes, post-HS	0.051	0.294	0.965	0.966	0.377	0.956	0.289	0.060	0.067	0.875
Ages 25+, Injuries, post-HS	0.993	0.957	0.881	0.813	0.213	0.641	0.758	0.648	0.628	0.499
Ages 5–9, HS-targeted causes, pre-HS	0.834	0.935	0.582	0.608	0.024	0.394	0.488	0.283	0.362	0.731
Whites, Ages 5–9, HS-targeted causes, post-HS	0.955	0.492	0.226	0.157	0.411	0.264	0.763	0.569	0.446	0.085
Blacks, Ages 5–9, HS-targeted causes, post-HS	0.127	0.058	0.090	0.105	0.307	0.764	0.666	0.410	0.327	0.155

Notes: (i) Point estimator is constructed using difference in means of unadjusted and adjusted outcomes, respectively, with a uniform kernel; (ii) Fisher's p-values are obtained by taking 10,000 draws from the distribution of the treatment assignment assuming a fixed-margins randomization; (iii) $\hat{h}_{\mathrm{LR}}$ corresponds to the bandwidth obtained by the method described in Cattaneo et al. (2015). Results were obtained using the implementations described in Cattaneo et al. (2016).

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A11. Comparison of inference approaches.

	Bandwidth/Window		Inference Results		
	Method	h	Method	RD TE	95% CI
Local Randomization					
$p = 0$	Covariate Balance	1.100	Finite Sample/Neyman/Fisher	−2.280	[−3.975, −0.575]
$p = 1$	Covariate Balance	1.100	Finite Sample/Neyman/Fisher	−2.515	[−4.225, −0.800]
Nonparametric Loc. Poly.					
$p = 0$	MSE-optimal	3.235	Asymptotic/Robust BC	−2.114	[−4.963, −0.149]
$p = 1$	MSE-optimal	6.811	Asymptotic/Robust BC	−2.409	[−5.462, −0.099]
Flexible Parametric					
$p = 1$	Ad-hoc	9.000	Asymptotic/Parametric	−1.895	[−3.828, 0.038]
$p = 1$	Ad-hoc	18.000	Asymptotic/Parametric	−1.198	[−2.498, 0.101]
Global Parametric					
$p = 4$	Full Sample	∞	Asymptotic/Parametric	−3.065	[−5.189, −0.940]

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Assumption 4 restricts the values of the treatment status to two, $d_i(1)$ and $d_i(0)$, and the values of the potential outcomes to two, namely, $\tilde{y}_i(0)$, and $\tilde{y}_i(1)$.

Under the null hypothesis that $\tau = \tau_0$, the adjusted responses $\tilde{Y}_i - \tau_0 D_i$ are fixed and unrelated to the instrument (Imbens & Rosenbaum, 2005). As long as the distribution of the instrument is known, so is the distribution of any statistic $T(\mathbf{Z}_{W_0}, \tilde{\mathbf{Y}}_{W_0} - \tau_0 \mathbf{D}_{W_0})$ and hence exact p-values can be obtained from the randomization distribution as discussed previously.

A possible statistic for hypothesis testing is the difference in means for units with $Z_i = 1$ and $Z_i = 0$:

$$T_{\texttt{AR}} = \frac{\sum_{i \in \mathscr{I}_0} (\tilde{Y}_i - \tau_0 D_i) Z_i}{\sum_{i \in \mathscr{I}_0} Z_i} - \frac{\sum_{i \in \mathscr{I}_0} (\tilde{Y}_i - \tau_0 D_i)(1 - Z_i)}{\sum_{i \in \mathscr{I}_0} (1 - Z_i)}$$

where, as before, $\mathscr{I}_0 = \{i : R_i \in W_0\}$. This statistic corresponds to the difference in the intercepts from two (reduced-form) regressions of $Y_i - \tau_0 T_i$ on the score above and below the cutoff. We label it `AR` for Anderson-Rubin, since even though it is not technically the Anderson-Rubin statistic, it captures the idea of using the reduced form coefficients. In practice, this estimator can be obtained by running a regression of $Y_i - \tau_0 D_i$ on Z_i, R_i and an interaction term, with $T_{\texttt{AR}}$ simply being the coefficient corresponding to Z_i. See Cattaneo et al. (2016) for further implementation details.

REFERENCES

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2017). Coverage error optimal confidence intervals for regression discontinuity designs. Working paper, University of Michigan.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2016). Regression discontinuity designs using covariates. Working paper, University of Michigan.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression discontinuity designs. Stata Journal, forthcoming.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82, 2295–2326.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. Journal of the American Statistical Association, 110, 1753–1769.

Cattaneo, M. D., Frandsen, B., & 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, 1–24.

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2016). Inference in regression discontinuity designs under local randomization. Stata Journal, 16, 331–367.

Imbens, G. W., & Rubin, D. B. (2015). Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. Journal of Econometrics, 142, 675–697.

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

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. Journal of the American Statistical Association, 102, 191-200.

Sekhon, J., & Titiunik, R. (2016). Understanding regression discontinuity designs as observational studies. Observational Studies, 2, 173–181.

Sekhon, J., & Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In M. D. Cattaneo & J. C. Escanciano (Eds.), Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38). Emerald Group Publishing, forthcoming.