

Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods

Vivian C. Wong

Northwestern University

Peter M. Steiner

University of Wisconsin

Thomas D. Cook

Northwestern University

In a traditional regression-discontinuity design (RDD), units are assigned to treatment on the basis of a cutoff score and a continuous assignment variable. The treatment effect is measured at a single cutoff location along the assignment variable. This article introduces the multivariate regression-discontinuity design (MRDD), where multiple assignment variables and cutoffs may be used for treatment assignment. For an MRDD with two assignment variables, we show that the frontier average treatment effect can be decomposed into a weighted average of two univariate RDD effects. The article discusses four methods for estimating MRDD treatment effects and compares their relative performance in a Monte Carlo simulation study under different scenarios.

Keywords: *regression-discontinuity; causal inference; evaluation*

Introduction

In a traditional regression-discontinuity design (RDD), units are assigned to treatment and comparison conditions solely on the basis of a single cutoff score on a continuous assignment variable. The discontinuity in the functional form of the outcome at the cutoff represents the treatment effect, or the average treatment effect at the cutoff. Proven by Goldberger (2008) in 1972, the design has been shown empirically to produce effects akin to an experiment's (Aiken, West, Schwalm, Carroll, & Hsuing, 1998; Berk, Barnes, Ahlman, & Kurtz, 2010; Black, Galdo, & Smith, 2007; Buddelmeyer & Skoufias, 2004; Shadish, Galindo, Wong, Steiner, & Cook, 2011). It has been used for program evaluations in

criminal justice (Berk & Rauma, 1983), medicine (Finkelstein, Levin, & Robbins, 1996), economics (see Lee & Lemieux, 2010, for review), and education (Gill, Lockwood, Martorell, Setodji, & Booker, 2007; Jacob & Lefgren, 2004a, 2004b; Wong, Cook, Barnett, & Jung, 2008).

However, units are often assigned to treatment on more than one continuous assignment variable. Recent applications of RDDs in education have had multiple assignment variables and cutoff scores available for treatment assignment. For example, Jacob and Lefgren (2004a) and Matsudaira (2008) examined the effects of summer remedial education programs that were assigned to students based on missing a reading score cutoff, a math cutoff, or both. Kane (2003) and van der Klaauw (2002) evaluated the effects of college financial aid offers on students' postsecondary school attendance using measures of income, assets, and grade point average (Kane, 2003) or grade point average and SAT scores (van der Klaauw, 2002) as multiple assignment variables in an RDD. Papay, Murnane, and Willett (2010) and Martorell (2004) looked at the effects of failing high school exit exams in two subject areas—English language arts and math—on the probability of students' graduating from high school. Finally, Gill et al. (2007) examined the effects of schools' failure to make Adequate Yearly Progress under No Child Left Behind by missing 1 of the 39 possible assignment criteria. All are examples of the multivariate regression-discontinuity design (MRDD), where treatment assignment is based on cutoffs for two or more covariates rather than a single point along an assignment variable. MRDDs are not unique to education; they also occur with increasing frequency in other fields of research, such as in the evaluation of labor market programs (Card, Chetty & Weber, 2007; Lalive, 2008; Lalive, Van Ours & Zweimüller, 2006).

However, an RDD with multiple assignment variables raises challenges that are distinct from those identified in a traditional RDD. Treatment effects for an RDD with multiple assignment variables may be identified across multiple cutoff frontiers as opposed to a single point along the assignment variable. Thus, analytic procedures for estimating treatment effects across a multidimensional space are more complex and require more observations than approaches for estimating a treatment effect at a single point along the assignment variable. Although Cook et al. (2009), Reardon and Robinson (2012), and Papay, Murnane, and Willett (2011) outline various procedures for estimating treatment effects in an MRDD, the proposed approaches have not been derived formally, nor have they been tested empirically to examine their relative benefits and disadvantages.

This article has three purposes. The first is to use potential outcomes notation (Holland, 1986; Rubin, 1974) to define the causal estimand τ_{MRD} for an MRDD with two assignment variables (M and R) and cutoffs. We show that the frontier average treatment effect τ_{MRD} may be decomposed into a weighted average of two univariate RDD effects, τ_M at the M -cutoff and τ_R at the R -cutoff. We introduce the term frontier average treatment effect to

emphasize that the MRDD estimates treatment effects only for the subpopulation of units located *at the cutoff frontier*, as opposed to the average treatment effect for the overall study population. This is analogous to the univariate RDD, where only the average treatment effect *at the cutoff* is estimated. In both cases, the average treatment effect of the study population may be inferred from the local estimates at the cutoff frontiers only when constant treatment effects can be reasonably assumed.

The second purpose of this article is to provide guidance on the complexities of choosing an appropriate causal estimand of interest. Because each frontier produces a separate impact estimate, treatment effects may be reported individually (τ_M and τ_R) or pooled across multiple frontiers (τ_{MRD}). We show that in most cases, the frontier-specific effects will be the preferred causal estimand over the frontier average treatment effect τ_{MRD} because the latter is not scale-invariant. That is, τ_{MRD} depends crucially on the metric and scaling of the assignment variables. Estimating τ_{MRD} makes sense only if either the frontier-specific treatment effects are homogeneous or the assignment variables' metrics and scales are comparable. In this article, we elaborate further on issues related to choosing an appropriate causal estimand in MRDDs, and highlight the contexts and conditions required for preferring frontier-specific effects over a pooled effect.

Finally, this article seeks to test four analytic approaches for estimating treatment effects in an MRDD—the frontier, centering, univariate, and instrumental variable (IV) approaches—and to identify the causal estimand/estimands produced by each approach. The frontier approach estimates treatment effects by first modeling the discontinuity at the cutoff frontier using parametric, semiparametric, or nonparametric procedures, and then by applying appropriate treatment weights to each cutoff frontier to estimate τ_{MRD} . The approach estimates the frontier average treatment effect (τ_{MRD}) and frontier-specific effects (τ_M and τ_R) simultaneously. It is a more flexible extension of an approach introduced by Berk and de Leeuw (1999), which relied on parametric regression estimation of the entire response surface under the assumptions of constant treatment effects and a correctly specified regression model. The frontier approach we propose relaxes these assumptions by allowing for heterogeneous treatment effects along the cutoff frontier. Its limitation, however, is that it estimates the frontier average treatment effect, as opposed to the more general average treatment effect estimated by Berk and de Leeuw's method. In the centering approach, all assignment variables are centered at their respective cutoffs, and each unit is assigned its minimum centered assignment score. The minimum assignment score is used then as the single assignment variable in a traditional univariate RDD to estimate τ_{MRD} . This approach was employed by Gill et al. (2007) in their evaluation of No Child Left Behind. In the univariate approach, researchers choose a single assignment variable and cutoff to estimate a frontier-specific effect, and exclude all observations that are assigned to treatment via the second assignment variable and cutoff. Jacob and Lefgren (2004a) applied this approach in their evaluation of Chicago

remedial education programs. Finally, in the IV approach, researchers use at least one assignment mechanism as an instrument for treatment receipt and designate units assigned by the second assignment variable and cutoff as treatment-misallocated cases. Cook et al. (2009) and Reardon and Robinson (in press) propose this approach for analyzing MRDDs, but it has yet to be examined empirically. For each approach, we discuss the causal quantities, theoretical underpinnings, and required assumptions. Through Monte Carlo simulations, we examine the performance of the four approaches. Overall, we find that the frontier, centering, univariate, and IV approaches succeed in producing unbiased treatment effects when their design and analytic assumptions are met.

Before introducing the MRDD, we give a brief description of the traditional RDD with a single assignment variable and cutoff (Hahn, Todd, & van der Klaauw, 2001; Trochim, 1984). We then show the MRDD as an extension of the traditional RDD, except that treatment effects are estimated for cutoff frontiers, as opposed to at a single cutoff point. Although we discuss MRDDs with two assignment variables only, the concepts and analytic approaches presented here extend to MRDDs with more than two assignment variables.

The RDD

The RDD with a Single Assignment Variable

Under the standard Rubin causal model, the causal treatment effect is estimated for a binary treatment intervention D , where $D_i = 0$ if unit i belongs to the control condition and $D_i = 1$ if it belongs to the treatment condition. Let $Y_i(0)$ and $Y_i(1)$ denote the pair of potential outcomes with $Y_i(0)$ as the potential control outcome, which is observed if unit i is not exposed to treatment ($D_i = 0$), and $Y_i(1)$ the potential treatment outcome which is observed if unit i is exposed to treatment ($D_i = 1$). In practice, we do not observe both potential outcomes for each unit (the fundamental problem of causal inference; Holland, 1986). Rather, we observe only the potential treatment or control outcome for each unit depending on the treatment received. Hence, the observed outcome can be written as a function of the potential outcomes: $Y_i = (1 - D_i) \cdot Y_i(0) + D_i Y_i(1)$.

In a traditional RDD, units are assigned to treatment solely on the basis of a cutoff score (z_c) on a continuous assignment variable (Z). The assignment variable is any measure taken prior to the treatment intervention. Let us assume that a unit gets assigned to the treatment condition if its scores below the cutoff of the assignment variable and to the control condition if its score is equal to or above the assignment variable: $D_i = 1$ if $Z_i < z_c$ and $D_i = 0$ if $Z_i \geq z_c$. When the assignment rule is implemented perfectly, the probability of receiving treatment drops at the cutoff from 1 to 0. More formally, the discontinuity in the probability of treatment receipt at the cutoff is $\lim_{z \downarrow z_c} E[D_i | Z_i = z] - \lim_{z \uparrow z_c} E[D_i | Z_i = z] = 1$. Such an RDD is called “sharp” as opposed to a “fuzzy” design (Trochim, 1984)

where due to noncompliance, the probability of treatment receipt does not switch from 1 to 0 at the cutoff, but exhibits a jump less than 1.

In a sharp RDD, the causal quantity of interest is the difference in the potential outcomes at the cutoff, that is,

$$\tau_{RD} = E[Y_i(1) - Y_i(0)|Z_i = z_c] = E[Y_i(1)|Z_i = z_c] - E[Y_i(0)|Z_i = z_c].$$

Note that τ_{RD} is the average treatment effect *at the cutoff*. Since we never observe control and treatment cases at the cutoff, the causal estimand is better defined in terms of the difference in limits of conditional expectations as we approach the cutoff from below and above:

$$\begin{aligned}\tau_{RD} &= \lim_{z \uparrow z_c} E[Y_i(1)|Z_i = z] - \lim_{z \downarrow z_c} E[Y_i(0)|Z_i = z] \\ &= \lim_{z \uparrow z_c} E[Y_i|Z_i = z] - \lim_{z \downarrow z_c} E[Y_i|Z_i = z].\end{aligned}\tag{1}$$

The second equality is with observed instead of potential outcomes. This holds because we observe only the potential treatment outcomes below the cutoff, and only the potential control outcomes above or at the cutoff. The difference in limits represents the discontinuity (treatment effect) at the cutoff. However, for τ_{RD} to be interpreted as a causal effect at the cutoff, the potential outcomes must be continuous at the cutoff (Hahn et al., 2001; Imbens & Lemieux, 2007; Lee & Lemieux, 2010):

$$\begin{aligned}\lim_{z \downarrow z_c} E[Y_i(0)|Z_i = z] &= \lim_{z \downarrow z_c} E[Y_i(0)|Z_i = z] \text{ and} \\ \lim_{z \uparrow z_c} E[Y_i(1)|Z_i = z] &= \lim_{z \uparrow z_c} E[Y_i(1)|Z_i = z].\end{aligned}$$

Later in this article, when we introduce the IV approach for estimating frontier-specific effects, we discuss the required assumptions and analytic procedure for estimating treatment effects in a “fuzzy” RDD.

The MRDD With Two Assignment Variables

The MRDD has an assignment process that is based on two or more assignment variables. In this article, we consider only sharp MRDDs with two assignment variables, R and M , with respective cutoffs r_c and m_c . Units are assigned to treatment if they miss cutoff r_c , m_c , or both. Figure 1 shows that units are assigned to the control condition C if they score above both cutoffs ($R_i \geq r_c, M_i \geq m_c$) and to the treatment condition T if they score below either cutoff ($R_i < r_c$ or $M_i < m_c$). We partition the treatment assignment space into three subsets: T_1 if units miss only cutoff r_c , T_3 if they miss only cutoff m_c , and T_2 if they miss both cutoffs. Though we partition the treatment space into three subspaces, we assume that all cases receive exactly the same treatment (otherwise, more than one potential treatment outcome needs to be

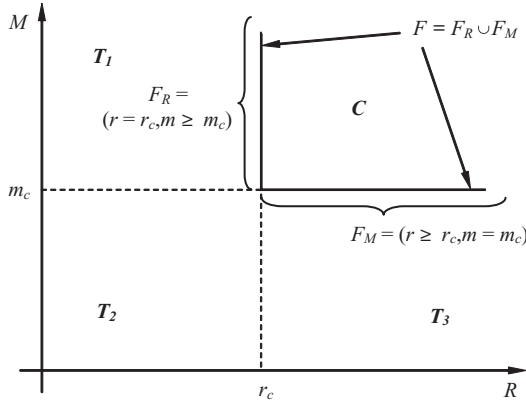


FIGURE 1. Multivariate Regression-Discontinuity Design (MRDD) with two assignment variables R and M .

considered). In this design, R and M may be reading and math test scores (respectively), treatment may be a standardized test preparation course, and assignment to treatment may be based on whether students fail to achieve minimum threshold scores for reading or math. Although this is a fairly specific implementation of an MRDD, the results presented here also apply to MRDDs where treatment and control conditions are swapped.¹ Figure 1 shows the cut-off frontier $F = \{(r, m) : (r \geq r_c, m = m_c) \cup (r = r_c, m \geq m_c)\}$ at which the frontier average treatment effect is estimated. Assuming complete treatment compliance, the frontier average treatment effect τ_{MRD} is given by the expected difference in potential outcomes at the cutoff frontier:

$$\tau_{\text{MRD}} = E[Y_i(1) - Y_i(0) | (R_i, M_i) \in F]. \quad (2)$$

Decomposition of the frontier average treatment effect τ_{MRD} . Since the cutoff frontier consists of the R -frontier along assignment variable M , $F_R = \{(r, m) : (r = r_c, m \geq m_c)\}$, and the M -frontier along assignment variable R , $F_M = \{(r, m) : (r \geq r_c, m = m_c)\}$, we can decompose the frontier average treatment effect into a weighted average of the treatment effects at the R - and M -frontiers. Let the difference in potential outcomes be $G_i = Y_i(1) - Y_i(0)$ and the joint density function for assignment variables R and M be $f(r, m)$, then, we can define the treatment effect at the cutoff frontier F as the weighted average of conditional expectations, given the single frontiers F_R and F_M (see Appendix A for the proof):

$$\begin{aligned} \tau_{\text{MRD}} &= E[G_i | (R_i, M_i) \in F] = w_R E[G_i | R_i \in F_R] + w_M E[G_i | M_i \in F_M] \\ &= w_R \tau_R + w_M \tau_M, \end{aligned} \quad (3)$$

where weights w_R and w_M reflect the probabilities for observing a unit at the R - or M -frontier,

$$w_R = \frac{\int_{m \geq m_c} f(r = r_c, m) dm}{\int_{m \geq m_c} f(r = r_c, m) dm + \int_{r \geq r_c} f(r, m = m_c) dr} \text{ and} \quad (4)$$

$$w_M = \frac{\int_{r \geq r_c} f(r, m = m_c) dr}{\int_{m \geq m_c} f(r = r_c, m) dm + \int_{r \geq r_c} f(r, m = m_c) dr}.$$

The conditional expectations represent the treatment effects τ_R and τ_M at the two discontinuity frontiers F_R and F_M since

$$\tau_M = E[G_i | M_i \in F_M] = \frac{\int_{r \geq r_c} g(r, m) f(r, m = m_c) dr}{\int_{r \geq r_c} f(r, m = m_c) dr} \text{ and} \quad (5)$$

$$\tau_R = E[G_i | R_i \in F_R] = \frac{\int_{m \geq m_c} g(r, m) f(r = r_c, m) dm}{\int_{m \geq m_c} f(r = r_c, m) dm},$$

where $g(r, m) = y_1(r, m) - y_0(r, m)$ is the difference in potential outcomes. Note that τ_R is the average treatment effect *at the R cutoff*, and τ_M is the average treatment effect *at the M cutoff*.

As shown in the Appendix A, we may also use conditional and marginal distributions for defining weights and conditional expectations, which is more convenient for estimating the treatment effects.

Assumptions required for MRDD. Given that the frontier average treatment effect may be decomposed into a weighted average of frontier-specific effects, all required assumptions for the traditional univariate RDD must be met for each discontinuity frontier (F_R and F_M). Thus, the design has two requirements: First, there must be a discontinuity in the treatment probabilities at F_R and F_M . Second, the expectations of potential outcomes must be continuous at F_R and F_M , but not necessarily along frontier F . The latter implies that meeting the continuity assumption for each cutoff frontier separately is sufficient for producing

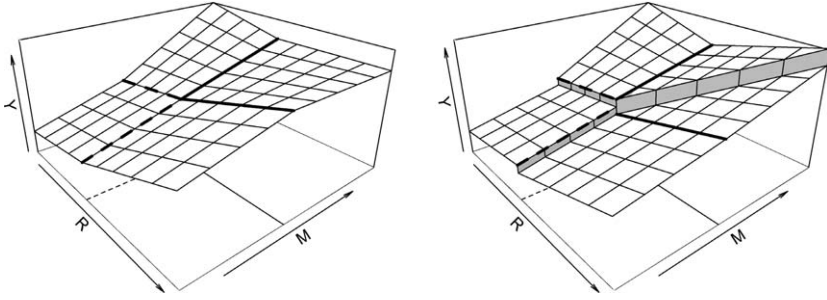


FIGURE 2. Response surface for potential treatment outcome $Y(1)$ without discontinuity (left panel) and with discontinuities (right panel).

unbiased results in the overall MRDD. We elaborate on this point below, and demonstrate our argument through the simulation study presented later in this article.

For the potential treatment outcomes, the continuity assumption states that the limits of the expected values have to be identical at the cutoff frontiers:

$$\lim_{r \downarrow r_c} E[Y_i(1)|R_i = r, M_i \geq m_c] = \lim_{r \downarrow r_c} E[Y_i(1)|R_i = r, M_i \geq m_c] \text{ and} \\ \lim_{m \downarrow m_c} E[Y_i(1)|R_i \geq r_c, M_i = m] = \lim_{m \downarrow m_c} E[Y_i(1)|R_i \geq r_c, M_i = m].$$

The same equality must hold for potential control outcomes $Y(0)$. Two important remarks need to be made. First, no continuity is required for $M < m_c$ at $R = r_c$ and for $R < r_c$ at $M = m_c$ because these frontiers do not belong to the discontinuity frontier F (in Figure 1, these are the dashed frontiers between treatment subsets T_1 and T_2 , and T_2 and T_3). Second, the continuity assumption does not imply continuity of $E[Y(1)]$ and $E[Y(0)]$ along frontier F . Consider an arbitrary point $r^* \geq r_c$ at cutoff frontier F_M . Then, $\lim_{r \downarrow r^*} E[Y_i(1)|R_i = r, M_i = m_c]$ may differ from $\lim_{r \downarrow r^*} E[Y_i(1)|R_i = r, M_i = m_c]$. In particular, the expectation of potential outcomes may be discontinuous at the intersection point ($r = r_c, m = m_c$), where frontiers F_R and F_M meet: $\lim_{m \downarrow m_c} E[Y_i(1)|R_i = r_c, M_i = m]$ may differ from $\lim_{r \downarrow r_c} E[Y_i(1)|R_i = r, M_i = m_c]$, and both may differ from $E[Y_i(1)|R_i = r_c, M_i = m_c]$ at the intersection point. Such a discontinuity at the intersection point has no impact on the frontier-specific treatment effects τ_M and τ_R because the intersection point has a probability mass of zero with respect to the discontinuity frontiers F_M and F_R . The same holds for the potential control outcome $Y(0)$.

Nonetheless, continuity in expectations of both potential outcomes along frontier F , particularly at the intersection point (r_c, m_c), is desirable in practice as Figure 2 illustrates (for clarity we only show the response surface of $E[Y(1)]$).

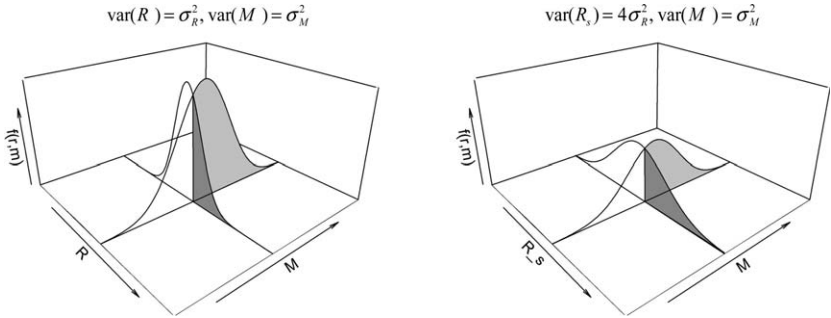


FIGURE 3. Scale-dependency of the joint density and weights at the cutoff frontier: By rescaling R such that $\text{var}(R_s) = 4\sigma_R^2$ the ratio of weights—represented by the ratio of the two areas along the frontier—changes. (A) $\text{var}(R) = \sigma_R^2$, $\text{var}(M) = \sigma_M^2$ (B) $\text{var}(R_s) = 4\sigma_R^2$, $\text{var}(M) = \sigma_M^2$.

The left panel shows a continuous response surface of $E[Y(1)]$ along the cutoff frontier F (solid line) but also along the entire cutoffs r_c and m_c (i.e., including the dashed lines). The right panel illustrates a case where $E[Y(1)]$ is discontinuous at the intersection point but continuous at each of the two frontiers F_M and F_R . This requires a rather awkward functional form with several discontinuities (e.g., a discontinuity along the diagonal for the control cases) to ensure that the potential treatment outcomes are connected smoothly at both frontiers. The response surface of the potential treatment outcome presented here seems rather implausible with a single treatment.

Scale-dependency of the frontier average treatment effect τ_{MRD} . The decomposition of the frontier average treatment effect of an MRDD into a weighted average of univariate RDD effects, τ_R and τ_M , reveals that the frontier average treatment effect τ_{MRD} depends on weights w_R and w_M . Since the weights are determined by integrating the joint density $f(r, m)$ along frontier F , their ratios depend crucially on the metric and scaling of assignment variables R and M . We use the terms “metric” to refer to the assignment variables’ units of measurement (e.g., income in thousand dollars, SAT score in points, or age in years), and “scaling” to refer to the spread (standard deviation) of the assignment variables, independent of whether they are on the same metric or not.²

To see how frontier weights are affected by the metric and scaling of the assignment variables, consider the following. The weight w_R for the treatment effect at the R -frontier decreases relative to w_M as assignment variable R is rescaled to $R_s = sR$ with $s > 1$ (with scaling of M held constant). Conversely, w_R ’s relative weight increases when the R assignment variable is rescaled with $s < 1$. This is because of the disproportional change in integrals $\int_{m > m_c} f(r = r_c, m) dm$ and $\int_{r > r_c} f(r, m = m_c) dr$ in Equation (4) when R and M are

rescaled with different scaling factors s . Figure 3 shows that when R is rescaled into $R_s = 2R$, such that the variance of the rescaled variable is four times larger than the original variable's variance, the ratio of the weights also changes (as indicated by the shaded areas along the cutoff frontier). Since the frontier average treatment effect τ_{MRD} is sensitive to the assignment variables' metric and choice of scale, an infinite number of average treatment effects exist. This is an unpleasant property of MRDD that is of special relevance whenever assignment variables are on a different metric or measurement scale and the treatment effects for frontiers F_M and F_R differ ($\tau_M \neq \tau_R$). We discuss practical implications of the scale-dependency and why standardizing assignment variables does little to address the problem in the discussion section.

Estimation Strategies for MRDD

An MRDD with two assignment variables allows the estimation of three different causal quantities: two frontier-specific effects, τ_R and τ_M , and the frontier average treatment effect τ_{MRD} . In this article, we present the following four estimation procedures: the frontier, centering, univariate, and IV approach. The first two approaches aim at estimating the frontier average treatment effect τ_{MRD} , and the latter two at the frontier-specific effects τ_R and τ_M . Table 1 provides a summary of the analytic procedure, the causal quantity estimated, the assumptions required, and the advantages and disadvantages of each procedure.

Frontier approach. This procedure estimates the discontinuity along both frontiers simultaneously and applies appropriate weights to obtain the frontier average treatment effect τ_{MRD} . It first estimates the discontinuous response surface $\hat{y}(r, m)$ using a parametric, semiparametric, or nonparametric regression method. Since we are interested only in the treatment effect at the cutoff frontier F , the treatment function $\hat{g}(r, m|F)$ is estimated by taking the difference in the estimated treatment outcome \hat{y}_1 and the control outcome \hat{y}_0 along discontinuity frontier F such that $\hat{g}(r, m|F) = \hat{y}_1(r, m|F) - \hat{y}_0(r, m|F)$. Then, the joint density function $\hat{f}(r, m)$ is estimated using a bivariate kernel-density estimator. Finally, we plug $\hat{g}(r, m|F)$ and $\hat{f}(r, m)$ into Equations (4) and (5) and estimate τ_R and τ_M , as well as weights w_M and w_R (for estimating τ_{MRD}), by numerical integration.

Because statistical software tools frequently do not offer procedures for estimating kernel-densities for three- or higher-dimensional data, nonparametric estimates of the univariate conditional and marginal densities $f_{R|M}$, $f_{M|R}$, f_R , and f_M might be used instead. By plugging the estimated univariate densities and treatment function into Equations (A3) and (A4) given in Appendix A and using numerical integration, we obtain frontier-specific effects $\hat{\tau}_R$ and $\hat{\tau}_M$, weights \hat{w}_R and \hat{w}_M , and the frontier average treatment effect $\hat{\tau}_{\text{MRD}}$. Because $\hat{f}(r, m)$ and marginal densities \hat{f}_R and \hat{f}_M depend on the metric and scaling of R and M , the choice of different bandwidths at the R - and M -cutoffs also would

TABLE 1
Summary of Analytic Approaches for Analyzing Sharp Multivariate Regression-Discontinuity Designs (MRDDs)

	Analytic Approaches			
	Frontier	Centering	Univariate	Instrumental Variable
Procedure	First, model the response surface using a parametric, semiparametric, or nonparametric method, and then use kernel-density estimates for integrating the treatment effects along the cutoff frontiers	Center all assignment variables at their respective cutoffs and assign each unit their minimum (or maximum) centered assignment score. Use the minimum assignment scores as the sole assignment variable and analyze as a traditional univariate RDD	Choose a single assignment variable and cutoff, excluding all observations that are assigned to treatment via the other assignment variables and cutoffs. Analyze with traditional univariate RDD methods	Use at least one assignment mechanism as an instrument for treatment receipt and designate units assigned by the other assignment mechanisms as fuzzy cases. Use standard methods for univariate fuzzy RDDs
Causal estimands	Frontier ATE (τ_{MRD}) and frontier-specific ATEs (τ_M, τ_R)	Only frontier ATE (τ_{MRD})	Only frontier-specific ATEs (τ_M, τ_R)	Only frontier-specific ATEs (τ_M, τ_R)
Design requirements	<ol style="list-style-type: none"> Discontinuity in treatment probabilities at each cutoff frontier. Continuity of potential outcomes at each cutoff frontier. Estimating τ_{MRD} requires that the assignment variables are on the same content and metric 	<ol style="list-style-type: none"> Discontinuity in treatment probabilities at each cutoff frontier. Continuity of potential outcomes at each cutoff frontier. For a meaningful interpretation, assignment variables need to be on the same metric in the same content area 	<ol style="list-style-type: none"> Discontinuity in treatment probability at the cutoff under investigation. Continuity of potential outcomes at the cutoff under investigation (i.e., the extended cutoff frontier). No defiers 	<ol style="list-style-type: none"> Discontinuity in treatment probability at the cutoff under investigation. Continuity of potential outcomes at the cutoff under investigation (i.e., the extended cutoff frontier). No defiers
Analytic requirements	Correct modeling of the entire response surface. Nonparametric methods require an optimal bandwidth choice and an appropriate degree of local polynomials. Kernel densities need to be reliably estimated using appropriate kernels and bandwidths	Correct modeling of the bivariate outcome regression. Nonparametric methods require an optimal bandwidth choice and an appropriate degree of local polynomials	Correct modeling of the bivariate outcome regression. Nonparametric methods require an optimal bandwidth choice and an appropriate degree of local polynomials	Correct modeling of the bivariate outcome regression and reliable estimation of the treatment compliance rate (the latter is guaranteed in a sharp MRDD). Nonparametric methods require an optimal bandwidth choice and an appropriate degree of local polynomials

(continued)

TABLE 1 (continued)

Analytic Approaches				
	Frontier	Centering	Univariate	Instrumental Variable
Advantages	Estimates all effects (τ_{MRD} , τ_M , and τ_R) simultaneously and, thus, is able to detect heterogeneous treatment effects. Higher efficiency (smaller standard errors) than centering approach	Handles MRDDs with many assignment variables with relative ease	Straightforward to compute and detects heterogeneous treatment effects. Higher efficiency (smaller standard errors) than the instrumental variable (IV) approach	None
Disadvantages	Relies on the strong assumption that the response surface and kernel density are correctly estimated. Nonparametric estimation of densities and numerical integration is cumbersome and computationally expensive (particularly if standard errors are bootstrapped).	Obscures heterogeneous treatment effects when they are present. More complex functional form of the centered assignment variable-outcome relationship, making treatment estimates prone to bias. Less efficient than frontier approach	Does not compute τ_{MRD}	No empirical test to assess whether continuity of potential outcomes for the extended frontier is valid. Less efficient than the univariate approach

Note. ATE, average treatment effect.

affect the ratio of weights. As a result, we recommend using the same bandwidth (in absolute units) for both dimensions. Bootstrapping may be used for estimating standard errors.

The advantage of the frontier approach is that it estimates the frontier average and frontier-specific effects simultaneously, detecting heterogeneous treatment effects if they are present. However, the approach requires a strong assumption that the response surface is correctly specified. In addition, the nonparametric estimation of densities and numerical integration is cumbersome, data-hungry, and computationally expensive (particularly in bootstrapping standard errors). The centering approach, which we discuss next, tries to overcome these issues by downscaling the multiple assignment variables into a single composite assignment variable.

Centering approach. This procedure collapses multiple assignment scores into a single assignment variable, thereby reducing a high-dimensional assignment mechanism to a one-dimensional mechanism. This is achieved by the following procedure: For each assignment variable, center each unit's score at its respective cutoff, such that $R_i^z = R_i - r_c$ and $M_i^z = M_i - m_c$. Then, choose the minimum centered value as the unit's sole assignment score: $Z_i = \min(R_i^z, M_i^z)$. The minimum applies only in MRDDs where the top right quadrant of the assignment variable plane is the control condition (quadrant C in Figure 1). In cases where a different segment of the surface is the control region (e.g., quadrant T2, see endnote 1), sign-transformations of the assignment variables or choosing the maximum centered assignment score is required for creating the composite assignment variable. Finally, apply standard RD analytic methods (e.g., local polynomial regression) for estimating treatment effects using Z as the assignment variable and zero as the cutoff.

Despite these dimension-reducing transformations, the centering approach estimates the same causal estimand as defined in Equation (2). First, note that at the cutoff z_c , where $\min(R_i^z, M_i^z) = 0$, the population consists of two subpopulations: units from the R -frontier $F_R = \{(r, m) : r - r_c = 0, m - m_c \geq 0\}$ and units from the M -frontier $F_M = \{(r, m) : r - r_c \geq 0, m - m_c = 0\}$. Then, the treatment effect can be decomposed further into the frontier average treatment effect as defined in Equation (3):

$$\begin{aligned} E[G_i|Z_i = 0] &= E[G_i|\min(R_i^z, M_i^z) = 0] \\ &= E[G_i|(R_i^z = 0, M_i^z \geq 0) \text{ or } (R_i^z \geq 0, M_i^z = 0)] \\ &= E[G_i|(R_i, M_i) \in F_R \cup F_M] = w_R E[G_i|R_i \in F_R] + w_M E[G_i|M_i \in F_M], \end{aligned}$$

where weights and conditional expectations are given as before (Equations (4) and (5)). This result implies that the causal quantity estimated by the centering approach is also sensitive to the metric and scaling of assignment variables. For example, increasing the variance for assignment variable R moves

observations for which $\min(R_i^z, M_i^z) = R_i^z$ farther away from the cutoff $z_c = 0$. Thus, units assigned by the R assignment mechanism would receive less weight relative to observations with $\min(R_i^z, M_i^z) = M_i^z$.

The chief advantage of the centering approach is that it allows the researcher to collapse scores from multiple assignment rules to a single assignment variable. The approach also generalizes well to MRDDs with more than two assignment variables, as well as simplifies the analyses for estimating the frontier average treatment effect across multiple discontinuity frontiers. However, collapsing the assignment variables into one score usually results in a more complex functional form, making the centering approach prone to misspecification bias even when nonparametric regression methods are used (a more detailed discussion is contained in the results section of the simulation study). Unlike the other three approaches, the centering approach does not estimate frontier-specific effects $\hat{\tau}_R$ and $\hat{\tau}_M$, but the frontier average treatment effect $\hat{\tau}_{MRD}$. Thus, this approach should be considered only when the assignment variables are comparable in terms of their metric units and scaling.

Univariate approach. This approach solves the dimensionality problem by estimating treatment effects for each frontier separately. We estimate the treatment effect at the R -frontier F_R by excluding all observations scoring on assignment variable M below the cutoff m_c since the cutoff frontier F_R is defined only for $M \geq m_c$. Then, using standard RDD methods like local polynomial regression (Imbens & Lemieux, 2007), we estimate the treatment effect at F_R according to Equation (1). We estimate the treatment effect at the cutoff frontier F_M in a similar way, except that we exclude observations scoring below the cutoff on assignment variable R . The frontier average treatment effect τ_{MRD} may be estimated, but it requires calculation of appropriate treatment weights for each frontier as described for the frontier approach. The univariate approach is most appropriate when separate treatment effects are desired for each frontier.

IV approach. Rather than excluding observations assigned to treatment by alternative mechanisms, we can estimate frontier-specific effects (for say, F_R) by including all units below and above the cutoff, and deriving an instrument for treatment receipt using the R -assignment variable and cutoff alone. Thus, all units for which $R_i \geq r_c$ and $M_i < m_c$ (units in quadrant T3 in Figure 1) are considered “noncompliers” of our designated assignment rule r_c . Although it may seem odd to treat cases in quadrant T3 as “noncompliers” in a sharp MRDD, the logic is similar to what would be applied in a univariate fuzzy RD, where the average complier effect at the cutoff is estimated using the assignment rule as an instrument for treatment receipt. In the MRD case, however, all units that are assigned to treatment based on alternative assignment rules (in this case, the M -cutoff m_c) are considered “fuzzy” units, while those assigned to treatment

based on our designated cutoff are “compliers.”³ If the continuity assumption is met, then the causal estimand identifies the treatment effect for the subpopulation of units that comply with treatment assignment at the cutoff (Imbens & Lemieux, 2007; Hahn *et al.*, 2001). More formally, unit i is a complier (C) if it adheres with treatment assignment $i \in C$ if $\lim_{r \uparrow r_i} D_i(r) = 1$ and $\lim_{r \downarrow r_i} D_i(r) = 0$, with D being the treatment indicator as defined earlier in the article. Here, unit i takes the treatment when assigned to treatment and it takes the control condition when assigned to the control. The causal estimand for τ_R at the R-frontier is then given by $\tau_R = E[Y_i(1) - Y_i(0)|R_i = r_c, i \in C]$ or in limits notation by

$$\begin{aligned} \tau_R &= \lim_{r \uparrow r_c} E[Y_i(1)|R_i = r, i \in C] - \lim_{r \downarrow r_c} E[Y_i(0)|R_i = r, i \in C] \\ &= \frac{\lim_{r \uparrow r_c} E[Y_i|R_i = r] - \lim_{r \downarrow r_c} E[Y_i|R_i = r]}{\lim_{r \uparrow r_c} E[D_i|R = r] - \lim_{r \downarrow r_c} E[D_i|R = r]}. \end{aligned} \quad (6)$$

Thus, the complier average treatment effect at r_c is given by the ratio of the difference in the new treatment and control group’s mean values and the difference in compliance rates at the cutoff. Nonparametric methods are typically used for estimating the treatment effect at the cutoff (Imbens & Lemieux, 2007).

In general, the complier average treatment effect at the cutoff r_c is a different causal quantity from the average treatment effect at the restricted cutoff frontier F_R . However, under the assumption of a sharp MRDD, only units with $M_i \geq m_c$ are considered as compliers while those with $M_i < m_c$ are not compliers because they always take treatment (i.e., they are assigned to treatment independent of the actual R assignment mechanism). Thus, the complier average treatment effect at the cutoff r_c is identical to the average treatment effect at the cutoff frontier F_R . It is important to note that the IV approach relies on a stronger assumption than the univariate approach because it requires continuity in the expected potential outcomes for all units at r_c to estimate the complier average treatment effect at cutoff r_c (as opposed to continuity at the restricted frontier F_R only). Given the stronger assumptions required—and the lack of empirical tests to probe these assumptions—we are reluctant to recommend the IV approach for applied practice. However, we include this approach in our empirical test because it has been proposed in the recently MRDD literature as a method for estimating treatment effects (Cook *et al.*, 2008; Reardon & Robinson, *in press*). So our purpose is to demonstrate the contexts and conditions when its assumptions are likely to be violated, and amount of bias that would occur in these cases.

Simulation Design for Monte Carlo Comparison of Approaches

In a series of simulation studies, we examine the performance of the frontier, centering, univariate, and IV approaches when we vary the following three

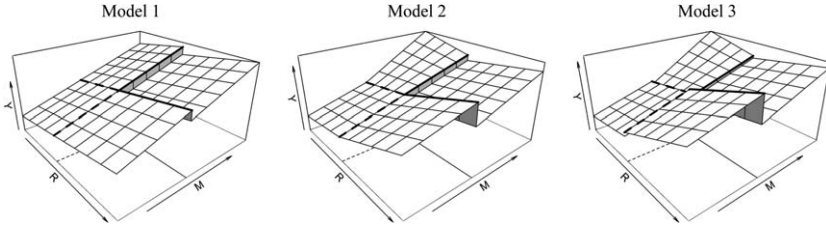


FIGURE 4. Response surfaces of observed outcomes for simulations: Constant treatment effects model (Model 1), heterogeneous treatment effects model (Model 2), and heterogeneous effects but with discontinuities in the potential treatment outcome model (Model 3).

factors: (a) complexity of the “true” response surface; (b) metric and scale of the assignment variables; and (c) approach for analyzing a sharp MRDD.

The first factor varies the complexity of the “true” response surface. The goal here is to assess how each approach performs when the “true” response surface is straightforward to model versus when it is complex with heterogeneous treatment effects. The outcome Y_i for unit i is a simulated math score based on three different specifications of the true response surface, and the assignment variables R and M are simulated reading and math test scores:

$$\text{Model 1: } Y_i = 4T_{0i} + .5R_i + 1M_i + \varepsilon_i,$$

$$\begin{aligned} \text{Model 2: } Y_i = & 4T_{0i} + .5R_i + 1M_i - .05T_{1i}M_i + .55T_{3i}R_i - .025T_{1i}R_iM_i \\ & - .005T_{3i}R_iM_i + \varepsilon_i, \end{aligned}$$

$$\begin{aligned} \text{Model 3: } Y_i = & 4T_{0i} - 2T_{1i} + 2T_{3i} + .5R_i + 1M_i - .05T_{1i}M_i + .55T_{3i}R_i \\ & - .025T_{1i}R_iM_i - .005T_{3i}R_iM_i + \varepsilon_i, \end{aligned}$$

where R_i and M_i are drawn from a bivariate normal distribution with a correlation of 0.2. T_{0i} equals 1 if unit i receives any treatment at all and 0 if it did not, and as depicted in Figure 1, T_{1i} equals 1 if unit i has an R assignment score less than r_c but an M assignment score greater than m_c (and 0 if otherwise), and T_{3i} equals 1 if unit i has an M assignment score less than m_c but an R assignment score greater than r_c (and 0 if otherwise). ε_i is a normally distributed error term with a mean of 0 and an SD of 2. Model 1 (constant treatment effects model) shows a constant treatment effect of 4.00, with no treatment by assignment variable interactions and level changes in the treatment response surface (Model 1 in Figure 4). Model 2 (heterogeneous treatment effects model) defines heterogeneous effects based on continuous potential treatment outcomes. In this model, effects along F_R and F_M are heterogeneous due to interactions between the treatment and assignment variables where students with different ability levels might

react differently to the same treatment (Model 2 in Figure 4). Our third model (heterogeneous but discontinuous effects model) defines heterogeneous effects based on discontinuous potential treatment outcomes. The model indicates heterogeneous treatment effects across the cutoff frontiers, as well as level changes in the treatment response surface between T_1 and T_2 and between T_2 and T_3 (Model 3 in Figure 4). Level changes in the response surface might occur if the treatment condition for T_1 and T_3 vary greatly from the treatment condition for T_2 . We include Model 3 to demonstrate the robustness of the four approaches when the “single treatment” assumption is violated, and when discontinuities in the response surface other than at the cutoff frontier are present. As discussed above, we expect that this model will produce biased-effect estimates for the IV approach.

Given the above data-generating equations and the distribution of assignment variables, we first computed for each model the true treatment effects for frontiers F_R and F_M and F . These theoretical effects serve as our benchmarks for comparing the estimates produced by the four approaches. Since the frontier average treatment effect τ_{MRD} depends on the metric and scaling of the assignment variables, we present two theoretical effects for each model: one for the raw, unstandardized assignment variables and one for the standardized assignment variables.

The second factor we vary examines how differences in the metric and scale of the two assignment variables in an MRDD affect the performances of the proposed approaches. First, we look at two assignment variables—say, reading and math test scores—that are on the same metric with identical distributional scales and shapes. Both variables R and M are normally distributed with the same SD of 10, but have different means and cutoffs. For the R assignment variable, the mean is 45 and the cutoff is 40. For the M assignment variable, the mean is 55 and the cutoff is 60. Second, we examine an MRDD with assignment variables that have the same means as above, but show differences in the distributions’ scale parameter, that is, their standard deviations differ. In this scenario, R has an SD of 5 and M has an SD of 20. Finally, we look at an MRDD with assignment variables on different metrics (e.g., SAT scores for reading and ACT scores for math). For the M assignment variable, we used the same mean, distribution, and cutoff as in the first scenario, but we transformed the R assignment variable such that $R_s = R/100$. Thus, the new R cutoff is .40, the mean is .45, and the SD is .10—this implies that the distribution of R has not only a different scale parameter but is also on a different metric (i.e., unit of measurement).

The third factor defines the four methodological procedures⁴ that we study. For the frontier approach, we investigate two parametric regression models with different specifications: the full model, which includes all covariates and interaction terms as defined in Model 3, and the constant treatment effects model, which assumes constant treatment effects across the response surface as in Model 1. This constant treatment effects model is

equivalent to Berk and de Leeuw's suggestion (1999).⁵ The goal is to assess the degree of bias when the treatment function is misspecified. We estimate the response surface via parametric regression and the conditional and marginal densities at the R - and M -cutoffs using a kernel-density estimator with an Epanechnikov kernel. Finally, we numerically integrate the product of treatment and density functions along the cutoff frontiers to obtain the conditional expectation across both frontiers (according to Equation (A4) in Appendix A).⁶

For the centering, univariate, and IV approaches, we estimate treatment effects using local linear kernel regression (Imbens & Lemieux, 2007). In each iteration of the simulation, the bandwidths are selected based on Imbens and Kalyanaraman's (2010) algorithm for optimal bandwidth choice at the cutoff. While the centering approach only allows the estimation of the frontier average treatment effect τ_{MRD} , the univariate and IV approaches focus on the average treatment effects at the cutoffs τ_M and τ_R . For the centering approach, we estimate treatment effects using both the raw and the standardized scores because recent applications of the centering approach (e.g., Gill et al., 2007) use standardized assignment variables for estimating treatment effects.

The goal of the Monte Carlo study with 500 simulated samples of size 5,000 is to evaluate the unbiasedness of the proposed approaches for estimating the true treatment effects. Due to computational reasons, we do not directly investigate variance estimators of the treatment effects (for each approach, standard errors may be bootstrapped). However, for assessing the relative efficiency of the four approaches, we report standard errors estimated from our simulation, which is the standard deviation of estimated treatment effects across the 500 iterations,

$$s_\tau = \sqrt{\frac{\sum_{i=1}^{500} (\hat{\tau}_i - \bar{\tau})^2}{500}}, \text{ where } \bar{\tau} \text{ is the average of the estimated effects } \hat{\tau}_i. \text{ Note}$$

that the investigation of the approaches' relative efficiency is restricted to a very specific simulation setting—we did not vary sample sizes, the location of cutoffs, and the assignment variables' correlation and distribution. To test the unbiasedness of an estimated treatment effect with respect to its corresponding true effect, we use simulation standard errors $s_\tau/\sqrt{500}$. Significant differences (at the .05 error level) between estimated treatment effects and true effects are indicated by asterisks in the tables (simulation standard errors are not presented in the tables).

Monte Carlo Results

Effect Estimates for MRDDs with Constant Treatment Effects

The first line of the column panels in Table 2 presents the theoretical treatment effects (τ_M , τ_R , and τ_{MRD}) according to the data-generating model (see Model 1, Figure 4). The second line shows the theoretical treatment effect

TABLE 2
Multivariate Regression-Discontinuity (MRD) Estimates for Model 1 (Constant Treatment Effects)

Assignment Variables (AVs)	Same Metric and Scale				Same Metric, Different Scales				Different Metrics			
	$R: \mu = 45, \sigma = 10, \text{Cutoff} = 40$ $M: \mu = 55, \sigma = 10, \text{Cutoff} = 60$				$R: \mu = 45, \sigma = 5, \text{Cutoff} = 40$ $M: \mu = 55, \sigma = 20, \text{Cutoff} = 60$				$R: \mu = .45, \sigma = .10, \text{Cutoff} = .40$ $M: \mu = .55, \sigma = .10, \text{Cutoff} = .60$			
Weights and effects	w_M	w_R	τ_M	τ_R	w_M	w_R	τ_M	τ_R	w_M	w_R	τ_M	τ_R
True effect: raw AVs	73.0	27.0	4.00	4.00	51.5	48.5	4.00	4.00	99.6	0.4	4.00	4.00
True effect: standard	73.0	27.0	4.00	4.00	80.9	19.1	4.00	4.00	73.0	27.0	4.00	4.00
Frontier approach: full model	72.8	27.2	3.99	4.01	51.3	48.7	3.99	4.02	4.01	99.6*	0.4*	4.01
Frontier approach: constant treatment effect	(3.1)	(3.1)	(0.1)	(0.19)	(4.5)	(4.5)	(0.1)	(0.23)	(0.13)	(0.1)	(0.1)	(0.20)
Centering approach: raw AVs	72.8	27.2	4.00	4.00	51.3	48.7	4.00	4.00	99.6*	0.4*	4.00	4.00
	(3.1)	(3.1)	(0.09)	(0.09)	(4.5)	(4.5)	(0.09)	(0.09)	(0.09)	(0.1)	(0.09)	(0.09)
Centering approach: standardized AVs					3.98				3.86*			4.02
					(0.47)				(0.90)			(0.28)
Univariate approach			3.98	3.99	3.94*				4.00			3.97*
			(0.50)	(1.02)	(0.35)				(0.50)			(0.36)
Instrumental variable (IV) approach	4.06	4.31					3.99	4.13			3.99	4.01
	(0.81)	(3.65)					(0.43)	(1.68)			(0.38)	(0.42)
							4.00	4.32			3.98	4.09
							(0.57)	(5.76)			(0.63)	(1.03)

Note. Weights w_M and w_R are given in percent ($w_M + w_R = 100$). Standard errors (estimated from the simulation) are in parenthesis.
*Indicates statistical difference ($p < .05$) between estimated and theoretical treatment effects (based on simulation standard errors).

when both assignment variables are standardized before analyzing the MRDD. Regardless of the metric and scale of the assignment variables, all the true effects are 4.00 when the raw or standardized assignment scores are used. This demonstrates that standardizing assignment scores does not affect the causal quantities estimated for the frontier-specific and frontier average treatment effects whenever these effects are constant. Note, however, that weights w_M and w_R remain sensitive to differences in the metric and scale of the assignment variables.

The first panel of Table 2 shows the results for two assignment variables on the same metric and scale with identical distributions (first panel of Table 2). The frontier approach results in treatment effects between 3.99 and 4.01 at the F_R and F_M frontiers, and 4.00 for the frontier average treatment effect. Of the four approaches examined, the frontier approach yields the most precise estimates. Treatment weights produced by the frontier approaches are also comparable in terms of relative proportion to the theoretical weights. The centering approach produces a frontier average effect of 3.98 when raw assignment scores are used and a frontier average effect of 3.94 when standardized scores are applied. The latter is significantly different from the theoretical effect, suggesting that it is slightly biased. Treatment effects' standard errors for both centering estimates are approximately four times larger than those generated by the frontier approach. The univariate approach replicates treatment effects for both frontiers, but again, standard errors are between 5 (F_M) and 10 times (F_R) larger than those produced by the frontier approach. Although the IV approach estimates effects that are not significantly different from the theoretical true effects, they are less efficient. The treatment standard error is .81 for τ_M and 3.64 for τ_R (the standard errors differ considerably due to the differential strength of instruments).

The second column panel of Table 2 shows weights and treatment effects when assignment variables are on the same metric but differ in the scale of their distributions (i.e., standard deviations). Except for the centering approach with unstandardized assignment variables, all methods replicate the theoretical treatment results. The frontier approach yields treatment estimates that range from 3.99 to 4.02, with treatment standard errors that are between .01 and .16. The centering approach yields a frontier average treatment effect of 3.86 when the raw assignment score is used, and an effect of 4.00 when standardized scores are applied. The treatment effect for the raw assignment score (3.86), however, slightly underestimates the theoretical effect of 4.00. The univariate and IV approaches produce effects that are not significantly different from their theoretical effects, but are less efficient than those generated by the frontier approach.

We also examined the performance of the proposed approaches when the MRDD is based on two assignment variables on different metrics (third column panel of Table 2). As before, all procedures replicate the theoretical effects of

4.00 points, except for the centering approach with standardized assignment variables. Estimates obtained by the frontier approach are the most efficient, while the IV estimates are the least efficient.

Overall, the frontier, univariate, and IV approaches succeed in replicating the theoretical estimates when treatment effects are constant. The exception is the centering approach, which produced biased results half the time. The significant, but small, bias is a result of pooling units from multiple cutoff frontiers into a single composite variable. This is because of two reasons (given nonconstant response surfaces for potential outcomes): First, pooling units from different frontiers increases the heterogeneity of the outcome at the pooled cutoff, requiring a larger bandwidth for nonparametric estimates. Second, pooling increases the complexity of the functional form around the cutoff. Even in the simple case with linear response surfaces and a constant treatment effect (Model 1), pooling produces a nonlinear relation between the composite assignment variable and the outcome due to differential distributions of frontier-specific units in the neighborhood of the single cutoff. The increased heterogeneity of the outcome and more complex functional form at the cutoff are likely to introduce bias in the local polynomial regression due to misspecification of the response function. Modeling a quadratic or cubic polynomial in the local regression would mitigate the bias, but it would also reduce the efficiency of the estimates. This result highlights the sensitivity of the centering approach to misspecifications in the response function, even when nonparametric regression methods are used. The univariate and IV approaches, which also use local linear regression for estimating effects, do not exhibit the same bias because the local linearity assumption holds for the simulated data.

Estimates for MRDDs With Heterogeneous Treatment Effects

Table 3 presents results for an MRDD with heterogeneous treatment effects across the treatment response surface (see Model 2, Figure 4). When the assignment variables are on the same metric with similar variances, the effects for F_M and F_R are 9.74 and 3.70, and for the frontier average treatment effect, it is 8.11 points (first two lines of the first panel in Table 3). Note that the frontier average effect is the same, regardless of whether the standardized or raw assignment scores are used, even when treatment effects are heterogeneous. When the MRDD consists of two assignment variables on the same metric but have different standard deviations, theoretical treatment estimates are 7.59 points for F_M and 3.36 points for F_R , for both the raw and the standardized effects (second panel). However, while the frontier-specific effects do not depend on scaling, the frontier average treatment effect does. When raw assignment scores are used, the theoretical frontier average effect is 5.54, but when standardized scores are used, the theoretical frontier effect is 6.79 points. The differences in these two effects

TABLE 3
Multivariate Regression-Discontinuity (MRD) Estimates for Model 2 (Heterogeneous Treatment Effects)

Assignment Variables (AVs)	Same Metric and Scale						Same Metric, Different Scales						Different Metrics					
	<i>R</i> : $\mu = 45$, $\sigma = 10$, Cutoff = 40						<i>R</i> : $\mu = 45$, $\sigma = 5$, Cutoff = 40						<i>R</i> : $\mu = .45$, $\sigma = .10$, Cutoff = .40					
	<i>M</i> : $\mu = 55$, $\sigma = 10$, Cutoff = 60						<i>M</i> : $\mu = 55$, $\sigma = 20$, Cutoff = 60						<i>M</i> : $\mu = 55$, $\sigma = 10$, Cutoff = 60					
Weights and Effects	w_M	w_R	τ_M	τ_R	τ_{MRD}	w_M	w_R	τ_M	τ_R	τ_{MRD}	w_M	w_R	τ_M	τ_R	τ_{MRD}	w_M	w_R	τ_{MRD}
True effect: raw AVs	73.0	27.0	9.74	3.70	8.11	51.5	48.5	7.59	3.36	5.54	99.6	0.4	9.74	4.00	9.72			
True effect: standard.	73.0	27.0	9.74	3.70	8.11	80.9	19.1	7.59	3.36	6.79	73.0	27.0	9.74	4.00	8.19			
Frontier approach: full model	72.5*	27.5*	9.77*	3.71	8.10	51.4	48.6	7.60	3.36	5.55	99.6	0.4	9.76	4.00	9.74			
	(3.1)	(3.1)	(0.28)	(0.20)	(0.31)	(4.5)	(4.5)	(0.17)	(0.24)	(0.26)	(0.1)	(0.1)	(0.27)	(0.19)	(0.27)			
Frontier approach: constant treatment effect	72.5*	27.5*	9.80*	9.80*	9.80*	51.4	48.6	7.57*	7.57*	7.57*	99.6	0.4	9.43*	9.43*	9.43*			
	(3.1)	(3.1)	(0.13)	(0.13)	(0.13)	(4.5)	(4.5)	(0.11)	(0.11)	(0.11)	(0.1)	(0.1)	(0.12)	(0.12)	(0.12)			
Centering approach: raw AVs					8.10					5.45*					9.75			
					(0.64)					(0.84)					(0.37)			
Centering approach: standardized AVs					8.07					6.80					8.12*			
					(0.52)					(0.48)					(0.52)			
Univariate approach			9.78	3.70				7.63	3.35				9.78	3.99				
			(0.73)	(0.97)				(0.54)	(1.51)				(0.55)	(0.40)				
Instrumental variable (IV) approach			9.87*	3.80				7.61	3.35				9.82	4.00				
			(1.24)	(3.71)				(0.74)	(5.61)				(0.92)	(0.97)				

Notes. Weights w_M and w_R are given in percent ($w_M + w_R = 100$). Standard errors (estimated from the simulation) are in parenthesis.
*Indicates statistical difference ($p < .05$) between estimated and theoretical treatment effects (based on simulation standard errors).

show that standardizing the assignment variables changes the relative proportion of the treatment weights when an MRDD with heterogeneous effects has two assignment variables with different distributions. Standardizing changes the weight ratio of frontiers F_M and F_R from 51.5:48.5 to 80.9:19.1 (Table 3 shows the weights in percentage terms that sum to 100). Finally, for an MRDD with assignment variables on different metrics and with heterogeneous treatment effects (third column panel of Table 3), the true frontier average treatment effect also depends on the assignment variables measurement scale due to the differences in weight ratios.

Table 3 shows that all four methods generally perform as expected when effects are heterogeneous along the cutoff frontiers. The frontier approach produces unbiased effects when the response function is correctly specified (full model), with the only exception being an estimate for F_M (first panel of Table 3). However, this exception is caused by the slight bias in the frontier weights, which is mostly likely due to chance because the frontier approach produces no other significant differences whenever the model is correctly specified. When the treatment functions are incorrectly modeled, the frontier approach performs poorly in reproducing the theoretical effects. The constant treatment effects model—which assumes constant effects across the treatment response surface—yields estimates that are significantly different from their benchmark effects (for raw assignment variables). This is true regardless of whether the assignment variables are on the same metric and scale, on the same metric with different distributional scales, or when the assignment scores are on different metrics. While the centering approach produces biased results in two of the six estimates (for the same reason as discussed above), the univariate and IV approaches produce unbiased effect estimates in every case. Also note that the centering approach with standardized assignment variables estimates the corresponding true effect for standardized scores (which differs from the true effect for raw scores).

*Estimates for MRDDs With Heterogeneous Treatment Effects
but Discontinuous Potential Treatment Outcomes*

Table 4 presents results for MRDDs with heterogeneous treatment effects but with discontinuous potential treatment outcomes (see Model 3, Figure 4). When the assignment variables are on the same metric and scale, the true treatment effect for F_M is 8.61 points, and for F_R , it is 1.70 points, making the frontier average treatment effect across both frontiers 6.74 points (first column panel). These effects are identical for when the raw and standardized assignment scores are used, showing again that when the assignment variables are on the same metric with identical standard deviations, standardizing does not change the relative proportions of the treatment weights. If the two

TABLE 4
Multivariate Regression-Discontinuity (MRD) Estimates for Model 3 (Heterogeneous Effects but With Discontinuities in the Potential Treatment Outcome)

Assignment Variables (AVs)	Same Metric and Scale				Same Metric, Different Scales				Different Metrics			
	R: $\mu = 45, \sigma = 10$, Cutoff = 40 M: $\mu = 55, \sigma = 10$, Cutoff = 60				R: $\mu = 45, \sigma = 5$, Cutoff = 40 M: $\mu = 55, \sigma = 20$, Cutoff = 60				R: $\mu = 45, \sigma = .10$, Cutoff = 40 M: $\mu = 55, \sigma = 10$, Cutoff = 60			
Weights and effects	w_M	w_R	τ_M	τ_R	w_M	w_R	τ_M	τ_R	w_M	w_R	τ_M	τ_R
True effect: raw AVs	73.0	27.0	8.61	1.70	6.74	48.5	7.63	1.36	4.59	99.6	0.4	8.61
True effect: standard	73.0	27.0	8.61	1.70	6.74	19.1	7.63	1.36	6.44	73.0	27.0	8.61
Frontier approach: full model	72.9	27.1	8.62	1.70	6.75	48.8	7.64	1.36	4.59	99.6	0.4	8.62
Frontier approach: constant treatment effect	(3.3)	(3.3)	(0.16)	(0.19)	(0.28)	(5.0)	(0.11)	(0.23)	(0.35)	(0.1)	(0.1)	(0.16)
	72.9	27.1	7.77*	7.77*	7.77*	48.8	6.83*	6.83*	6.83*	99.6	0.4	7.42*
	(3.3)	(3.3)	(0.11)	(0.11)	(0.11)	(5.0)	(0.10)	(0.10)	(0.10)	(0.1)	(0.1)	(0.11)
Centering approach: raw AVs					6.75				4.58			8.59
					(0.52)				(0.88)			(0.30)
Centering approach: standardized AVs					6.73				6.49*			6.77*
					(0.41)				(0.47)			(0.46)
Univariate approach			8.58	1.72			7.63	1.40			8.59	2.00
			(0.59)	(0.90)			(0.47)	(1.61)			(0.47)	(0.43)
Instrumental variable (IV) approach			9.42*	3.80*			7.96*	2.70*			9.33*	3.50*
			(1.22)	(3.11)			(0.84)	(5.10)			(0.89)	(1.72)

Notes. Weights w_M and w_R are given in percent ($w_M + w_R = 100$). Standard errors (estimated from the simulation) are in parenthesis.
*Indicates statistical difference ($p < .05$) between estimated and theoretical treatment effects (based on simulation standard errors).

assignment variables differ in their distributions' scale, the theoretical frontier average treatment effect for the raw assignment variables is 4.59 points but 6.44 points when assignment variables are standardized. The difference between the raw and standardized theoretical effect is reflected in the frontier weights, 51.5 and 48.5 for F_M and F_R , respectively, when raw scores are used, and 80.9 and 19.1 when standardized scores are applied. Similarly, for assignment variables on different metrics, the frontier average treatment effect depends on the scaling: 8.59 for the raw scores and 6.82 for the standardized scores (third panel of Table 4).

Simulation results presented in Table 4 show that when effects are heterogeneous but with discontinuities in potential treatment outcomes, the IV approach fails to generate unbiased-effect estimates for either frontier. The biases are large and significantly different from the theoretical benchmarks. The centering approach continues to produce mixed results, with small biases for a third of the estimates. The frontier approach generally performs well if the treatment function is correctly specified, and the univariate approach does well when the bandwidths and the degree of the polynomial are correctly specified for local nonparametric estimates.

Discussion

What do results presented in this article imply for practice? Our analytic and simulation work highlights the complexities of choosing an appropriate causal estimand in an MRDD. In many cases, the frontier average treatment effect may not have a meaningful interpretation because it does not make sense to pool effects across multiple frontiers. If at one frontier, the estimate indicates no effect and, at the other frontier, a significant positive effect, then the average effect across the entire frontier rests on a scale-dependent weighting scheme. In these cases, we recommend that researchers estimate frontier-specific effects because τ_M and τ_R can provide at least upper and lower bounds for the overall treatment effect. In addition, without strong assumptions (e.g., constant treatment effects), the frontier-specific effects τ_M and τ_R are less general than what would be obtained from a traditional univariate RDD with a corresponding assignment variable and cutoff. That is because the cutoff of a traditional RDD is not restricted by the cutoffs of additional assignment variables (e.g., units with $M_i < m_c$ are excluded for estimating treatment effects at F_R). Still, the presence of multiple cutoff frontiers has the advantage of exploring the heterogeneity of treatment effects along different dimensions. Finally, the frontier-specific and frontier average treatment effect cannot be generalized beyond the subpopulation of units that is close to the cutoff frontiers. As with standard RDD, MRDD produces only the treatment effects along the cutoff frontier/frontiers as opposed to across the entire response surface. Thus, researchers have the onus of

communicating to practitioners and policy makers which causal quantities are evaluated, explaining why these are the causal quantities of interest, and discussing the benefits and limitations of the results.

One common question raised about the frontier average treatment effect is whether standardizing raw assignment scores for the centering and frontier approaches can address τ_{MRD} 's sensitivity to scale and metric differences. Consider two possible assignment variables on different metrics: household income and age. Here, treatment weights depend heavily on the units of measurement for each assignment variable, where a measure of income in thousands of dollars instead of dollars would drastically increase the weight of the "income frontier" in the frontier average treatment effect. However, standardizing the income and age variables, such that both have an *SD* of 1, does not solve the scaling issue because the procedure fails to provide a substantive interpretation of the treatment effect at the cutoff frontier. Standardizing is only one possibility for transforming assignment scores into a common metric—another option is to use rank-order scores. But all these methods result in different weights for the treatment frontiers and, thus, different causal quantities. A second concern with standardizing assignment variables is that the procedure is sensitive to distribution properties that would have strong effects on the standard deviation used for standardizing. For example, the presence of extreme observations for one assignment variable would result in a down-weighting of the treatment effect for the corresponding frontier (due to the large standard deviation). The only case when standardization procedures or other transformations are not an issue for estimating τ_{MRD} is when the treatment effects for both frontiers are identical, though we believe this scenario to be rare in practice.

However, there are scenarios when τ_{MRD} is an appropriate—and even preferred—causal estimand. In MRDDs where the assignment variables are on the same metric in the same content area, it is reasonable to pool effect estimates across multiple cutoff frontiers. Examples include when occupants are assigned to treatment on the basis of geographic coordinates (e.g., intersecting streets); when patients are assigned to treatment if their blood pressure exceeds a certain threshold in all measurements; when unemployed persons are assigned to a labor market program if their inflation-adjusted incomes fall below a certain threshold over two consecutive years; and, when a cohort of students receive coaching if they fail a certain aptitude test 2 years in a row. Note that the aptitude test for both measurement periods has to be on the same metric. Two different reading tests (e.g., for different grades), or a reading and a math test are not on the same metric, even if their scores are calibrated to appear comparable. This is because the test scores' similarities are typically the result of deliberate standardization procedures. Thus, the frontier average treatment effect and its interpretation would depend on the calibration of test scores.

Given the challenges with choosing an appropriate causal estimand in the MRDD, our recommendation for practice is to start with analyzing each frontier of the MRDD separately using the univariate approach. Results from our simulation study indicate that the nonparametric univariate approach performed well in estimating frontier-specific effects, with no significant differences between the estimated and theoretical effects for F_M and F_R , given a reasonable choice of the local polynomials degree and bandwidths. After estimating the frontier-specific effects, the researcher can assess whether treatment effects are constant across both cutoff frontiers, and if so, use the frontier or centering approach to estimate an overall effect τ_{MRD} . In general, the frontier approach performs well in estimating τ_{MRD} when the treatment functions are correctly specified. It also has the advantage of improved statistical efficiency for both the frontier average and the frontier-specific effects. Its main limitation, however, is that it requires a correct modeling of the response surface and a reliable estimate of the multivariate density or univariate marginal and conditional densities. Because the true functional form of the response surface is hardly ever known in practice, researchers may consider using the centering approach for estimating τ_{MRD} to avoid the complications of modeling a multidimensional response surface and estimating densities. The issue here is that the centering approach may be prone to small biases due to wider bandwidths and more complex functional forms caused by pooling units from different cutoff frontiers. However, the bias might be mitigated using difference scores as the outcome whenever pretests are available to reduce the complexity of the functional form and heterogeneity of the dependent variable.

Overall, we recommend against using the IV approach for estimating frontier-specific effects τ_M and τ_R . Although it yields unbiased estimates when its analytic assumptions are met, our simulation results indicate that the IV approach has reduced statistical precision as compared to the other three methods. The approach also yields biased results when there are discontinuities along the extended cutoff frontiers in the potential treatment outcomes. Thus, within the context of our simulation setting, the IV approach appears to offer no comparative advantage over the other three proposed methods.

As we have shown, assessing whether treatment effects are constant across the different frontiers is critical for choosing an appropriate causal estimand in an MRDD. Thus, a practical question is how one might investigate the heterogeneity of treatment effects across and along the frontiers. Figure 5 provides an example of plots that show the treatment effect as a function of the assignment variables. The X -axes represent the assignment variables for different frontiers; the Y -axes show the estimated treatment effect as a function of the assignment variables. The overlaid kernel-density curves indicate the relative weight that treatment estimates along the assignment variable receive in the frontier-specific effects and the pooled-frontier average

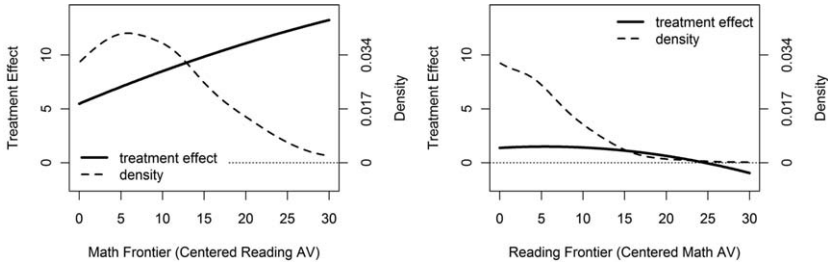


FIGURE 5. *Frontier-specific treatment effects with overlaid kernel-density estimates for the math and reading frontier estimated from a sample of Model 3.*

treatment effect. If treatment effects across the frontiers are constant, then estimating τ_{MRD} may be well warranted. In Figure 5, the first panel shows that the estimated treatment effect starts off at around 5 points near the origin and gets larger along the math frontier (or as the centered reading assignment scores increase). The density curve indicates that most units at the math frontier exhibit an average treatment effect between 5 and 10 points. The second panel of Figure 5 shows that the estimated treatment effect along the reading frontier is 2 points near the origin but gets slightly smaller and even becomes negative as the centered math assignment variable increases. However, according to the density plot, almost all units at the reading frontier have a treatment effect of around 2 points. Only a few units with a centered math assignment score of greater than 15 show smaller or negative effects. Overall, these plots indicate that treatment effects across the frontiers are heterogeneous, so pooling of frontier-specific effects is not well advised.

In this article, we have assumed that units are assigned to a single treatment condition via two assignment variables and cutoffs, but this need not be the case. Units may be assigned to one treatment condition for missing the R -cutoff, another if they miss the M -cutoff, and both if they miss the R - and M -cutoffs together. This type of MRDD raises several practical considerations for the researcher. First, although the frontier and univariate assignment variable approaches can estimate unbiased frontier-specific effects (τ_M and τ_R), the IV approach may yield biased results due to violations in the continuity assumption because variations in treatment conditions may introduce discontinuities in potential treatment outcomes. Second, an MRDD with multiple treatment conditions raises questions about whether estimating a frontier average treatment effect is appropriate and substantively interpretable, given that it averages estimates across two unique treatments. Third, treatment contrasts in an MRDD are limited to comparisons along the cutoff frontiers. There may be cases, however, when the desired treatment contrast is to compare outcomes from units that received the “strongest” treatment dosage with those that received no treatment at all. In our example, units in quadrant T_2 might receive

a stronger dosage of treatment because they missed both the R - and M -cutoffs (Figure 1). However, the MRDD does not include these observations in the calculation of treatment effects because they are not located at F_M or F_R . A researcher may choose to redefine observations in T_2 as the treatment units, and those in the remaining three quadrants as the comparison cases. However, this treatment contrast would involve comparing outcomes for units that missed both the R - and M -cutoffs (treatment cases) with those that missed only one cutoff (comparison cases). Again, this may not be the desired treatment contrast because treatment units are compared to those that have received at least some treatment because they missed either the M - or R -cutoffs. This predicament highlights one of the main design disadvantages of the MRDD, and one that researchers should be aware of as they interpret their treatment effects.

As with all Monte Carlo simulations, our study is limited by the fact that we cannot address every scenario that researchers are likely to encounter in analyzing MRDDs. First, we only examined the performance of the four approaches in the context of the sharp MRDD, where we assumed no instances of treatment misallocation. However, in many applications of MRDD, treatment crossover and no show are likely to occur. Traditionally, fuzziness around the RD cutoff is addressed using the assignment mechanism as an instrument for treatment receipt. This method extends to the MRDD for the univariate and centering approaches, where the treatment assignment mechanism again serves as an instrument for treatment receipt to estimate the local average treatment effect among compliers along the treatment frontier. Because of dimensionality issues, using an IV to address noncompliance for the frontier approach seems to be more challenging. A second limitation is that our study focuses on estimating treatment effects for MRDDs with only two assignment variables. Although the univariate, frontier, and centering approaches generalize well to MRDDs with more than two assignment variables, further work is needed for examining whether there are special analytic issues and requirements for analyzing MRDDs with more complex assignment mechanisms (e.g., with exemption rules like those in No Child Left Behind). Finally, our simulation does not examine how variations in the correlation and distribution of the assignment variables, the location of cutoffs, and the sample size would affect the relative efficiency of the four investigated methods. Overall, we expect sample size requirements for MRDDs to be higher than what is typically required for traditional univariate RDDs (which are already less efficient than comparable randomized experiments). The required sample size for the MRDD strongly depends on the distribution of the assignment variables and location of the cutoffs, as well as on the estimand of interest and the estimation approach used. In addition, one might be concerned about the efficiency of treatment effect estimates generated by the univariate approach when the correlation between the two

assignment variables is high, as may be the case in many education MRDDs where reading and math test scores are used as assignment variables. Further work that examines the challenges of using these four approaches in actual education contexts is needed.

Appendix

Proof of Decomposition

Let $g(r, m) = y_1(r, m) - y_0(r, m)$ be the difference in potential treatment and control outcomes and $f(r, m)$ the joint density of R and M . Then, using line integration along frontier F , with ds being an infinitesimal segment along the frontier F , the expected treatment effect given F can be decomposed into the weighted average of conditional expectations:

$$\begin{aligned}\tau_{\text{MRD}} = E[G|F] &= \frac{\int_F g \cdot f ds}{\int_F f ds} = \frac{\int_{F_R} g \cdot f ds + \int_{F_M} g \cdot f ds}{\int_{F_R} f ds + \int_{F_M} f ds} \\ &= \frac{\int_{F_R} f ds}{\int_{F_R} f ds + \int_{F_M} f ds} \cdot \frac{\int_{F_R} g \cdot f ds}{\int_{F_R} f ds} + \frac{\int_{F_M} f ds}{\int_{F_R} f ds + \int_{F_M} f ds} \cdot \frac{\int_{F_M} g \cdot f ds}{\int_{F_M} f ds} \\ &= w_R \cdot E[G|F_R] + w_M \cdot E[G|F_M].\end{aligned}$$

Since either R or M but never both assignment variables change simultaneously along the cutoff frontier, we can rewrite the line integrals defining weights and conditional expectations in terms of regular integrals along each of the assignment variables. We rewrite the treatment weights as

$$\begin{aligned}w_R &= \frac{\int_{m \geq m_c} f(r = r_c, m) dm}{\int_{m \geq m_c} f(r = r_c, m) dm + \int_{r \geq r_c} f(r, m = m_c) dr}, \\ w_M &= \frac{\int_{r \geq r_c} f(r, m = m_c) dr}{\int_{m \geq m_c} f(r = r_c, m) dm + \int_{r \geq r_c} f(r, m = m_c) dr}\end{aligned}\tag{A1}$$

and the conditional expectations as

$$\begin{aligned}
E[G|F_R] &= \frac{\int_{m \geq m_c} g(r, m) f(r = r_c, m) dm}{\int_{m \geq m_c} f(r = r_c, m) dm} \text{ and} \\
E[G|F_M] &= \frac{\int_{r \geq r_c} g(r, m) f(r, m = m_c) dr}{\int_{r \geq r_c} f(r, m = m_c) dr}.
\end{aligned} \tag{A2}$$

To simplify estimation of treatment weights and conditional expectations, it is useful to replace the bivariate density f by the product of the univariate conditional and marginal densities, $f = f_{R|M} \cdot f_M$ and $f = f_{M|R} \cdot f_R$, such that the treatment weights are given by

$$\begin{aligned}
w_R &= \frac{\int_{m \geq m_c} f_{M|R}(m|r = r_c) dm \cdot f_R(r = r_c)}{\int_{m \geq m_c} f_{M|R}(m|r = r_c) dm \cdot f_R(r = r_c) + \int_{r \geq r_c} f_{R|M}(r|m = m_c) dr \cdot f_M(m = m_c)} \text{ and} \\
w_M &= \frac{\int_{r \geq r_c} f_{R|M}(r|m = m_c) dr \cdot f_M(m = m_c)}{\int_{m \geq m_c} f_{M|R}(m|r = r_c) dm \cdot f_R(r = r_c) + \int_{r \geq r_c} f_{R|M}(r|m = m_c) dr \cdot f_M(m = m_c)}
\end{aligned} \tag{A3}$$

and conditional expectations by

$$\begin{aligned}
E[G|F_R] &= \frac{\int_{m \geq m_c} g(r = r_c, m) \cdot f_{M|R}(m|r = r_c) dm}{\int_{m \geq m_c} f_{M|R}(m|r = r_c) dm} \text{ and} \\
E[G|F_M] &= \frac{\int_{r \geq r_c} g(r, m = m_c) \cdot f_{R|M}(r|m = m_c) dr}{\int_{r \geq r_c} f_{R|M}(r|m = m_c) dr}.
\end{aligned} \tag{A4}$$

Acknowledgments

The authors thank J. R. Lockwood for helpful discussions, and J. D. Angrist and S. Reardon for their comments on an earlier version of the article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by grant R305U070003 and R305D100033 from the Institute for Educational Sciences, U.S. Department of Education.

Notes

1. Such a situation would occur if units were assigned to treatment if they made either cutoff, such as assignment to a gifted achievement program.
2. Note that metric and scale are conceptually distinct terms. For example, two assignment variables—such as income in thousands of dollars and age in years—may have the same mean and standard deviation but are on separate metrics because the units of measurement have different substantive interpretations.
3. The terminology usually used for the IV estimator—compliers, always-takers, never-taker, and defiers (Angrist, Imbens, & Rubin, 1996)—does not directly apply here because under the assumption of a sharp MRDD, all units comply with treatment according to the multiple assignment rules. In using the IV approach for estimating frontier-specific treatment effects, “noncompliers” are like always-takers because though they should be assigned the control condition based on our *designated* assignment rule, they take up treatment anyway because of the alternative assignment rules. Given a sharp MRDD, the monotonicity—or “no defiers”—assumption required for a consistent IV estimate is automatically met.
4. Stata code for the frontier, centering, univariate, and IV approaches are available upon request.
5. In using parametric regression, we could have estimated the frontier average and frontier-specific effects directly without integrating along the frontier (given a constant treatment effect, integration would have no effect on the average treatment effect at the cutoff frontier). However, a semiparametric or nonparametric estimation of the response surface would have required integration for estimating average treatment effects.
6. Although it was possible to estimate the bivariate kernel densities here, we used the univariate marginal and conditional densities for estimating frontier weights because we were interested in examining a procedure that easily generalizes to analyzing MRDDs with more than two assignment variables for

which multivariate kernel-estimators are not available in most statistical software packages.

References

- Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, 22, 207–244.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 87, 328–336.
- Berk, R. A., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, 6, 191–208.
- Berk, R. A., & de Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association*, 94, 1045–1052.
- Berk, R. A., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, 78, 21–27.
- Black, D., Galdo, J., & Smith, J. A. (2007). *Evaluating the regression discontinuity design using experimental data*. Retrieved from http://economics.uwo.ca/newsletter/misc/2009/smith_mar25.pdf
- Buddelmeyer, H., & Skoufias, E. (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA*. (World Bank Policy Research Working Paper No. 3386; IZA Discussion Paper No. 827). Retrieved from <http://ssrn.com/abstract=434600>
- Card, D., Chetty, R., & Weber, A. (2007). Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. *Quarterly Journal of Economics*, 122, 1511–1560.
- Cook, T. D., Wong, V. C., Steiner, P. M., Taylor, J., Gandhi, A., Kendziora, K., Choi, K., Garet, M., Jackson, S., Danielson, L., & Herman, R. (2009). *Impacts of school improvement status on students with disabilities: Feasibility report*. Washington, DC: American Institutes for Research.
- Finkelstein, M., Levin, B., & Robbins, H. (1996). Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *Journal of Public Health*, 86, 696–705.
- Gill, B., Lockwood, J. R., Martorell, F., Setodji, C. M., & Booker, K. (2007). *State and local implementation of the No Child Left Behind Act*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.
- Goldberger, A. S. (2008). Selection bias in evaluating treatment effects: Some formal illustrations. In T. Fomby, R. C. Hill, D. L. Millimet, and J. A. Smith, E. J (Eds.), *Modelling and evaluating treatment effects in economics* (pp. 131). Vytlacil, Amsterdam: JAI Press.
- Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69, 201–209.

- Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association*, 81, 945–960.
- Imbens, G. W., & Lemieux, T. (2007). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615–635.
- Imbens, G., & Kalyanaraman, K. (2010). *Optimal bandwidth choice for the regression discontinuity estimator*. Unpublished manuscript.
- Jacob, B., & Lefgren, L. (2004a). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, LXXXVI, 226–244.
- Jacob, B., & Lefgren, L. (2004b). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, 39, 50–79.
- Kane, T. J. (2003). A quasi-experimental estimate of the impact of financial aid on college-going. *Working Paper 9703*, National Bureau of Economic Research.
- Lalive, R. (2008). Unemployment benefits, unemployment duration, and post-unemployment jobs: A regression discontinuity approach. *American Economic Review*, 97, 108–112.
- Lalive, R., Van Ours, J. C., & Zweimüller, J. (2006). How changes in fiscal incentives affect the duration of unemployment. *Review of Economic Studies*, 73, 1009–1038.
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48, 281–355.
- Martorell, P. (2004). *Do high school graduation exams matter? A regression-discontinuity approach*. University of California, Berkeley.
- Matsudaira, J. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, 142, 829–850.
- Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, 32, 5–23.
- Papay, J. P., Murnane, R. J., & Willett, J. B. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161, 203–207.
- Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness* 1(5), 83–104.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Shadish, W. S., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods* 16, 179–191.
- Trochim, W. M. K. (1984). *Research design for program evaluation*. Beverly Hills, CA: Sage.
- van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43, 1249–1287.
- Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, 27, 122–154.

Authors

VIVIAN C. WONG is a postdoctoral fellow at the Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, Illinois 60208; email: vivianw@northwestern.edu. Her research interests include program evaluation in education contexts, causal inference, empirical tests of quasi-experimental designs, and implementation issues in regression-discontinuity design studies.

PETER M. STEINER is an assistant professor at the University of Wisconsin, Madison, Department of Educational Psychology, 1061 Educational Sciences, 1025 West Johnson Street, Madison, WI 53706; email: psteiner@wisc.edu. His research interests include causal inference, experimental and quasi-experimental designs, propensity score matching, and vignette experiments.

THOMAS D. COOK is the Joan and Sarepta Harrison Chair in Ethics and Justice, Professor of Sociology, Psychology, Education and Social Policy, and faculty fellow at the Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208; email: t-cook@northwestern.edu. His research interests include social science research methodology, program evaluation, whole school reform, and contextual factors that influence child and adolescent development.

Manuscript received August 16, 2010

First revision received February 1, 2011

Second revision received August 12, 2011

Accepted November 4, 2011