Running head: MULTIPLE RATING-SCORE REGRESSION DISCONTINUITY

Regression discontinuity designs with multiple rating-score variables

Sean F. Reardon
Stanford University

Joseph P. Robinson
University of Illinois at Urbana-Champaign

July 2010

An earlier version of this paper was presented at the Annual Meeting of the Society for Research on Educational Effectiveness, March 1-3, 2009. This research was supported by a grant from the Institute for Education Sciences (grant R305D100027), and benefitted from conversations with Howard Bloom, Marie-Andree Somers, and Michael Weiss. All errors remain our own.

Sean F. Reardon
Associate Professor of Education and (by courtesy) Sociology
School of Education, CERAS Building
520 Galvez Mall, #526
Stanford University
Stanford, CA 94305-3084
650.736.8517 (office phone)
650.723.9931 (office fax)
sean.reardon@stanford.edu

Joseph P. Robinson
Assistant Professor of Quantitative and Evaluative Research Methodologies
Department of Educational Psychology
University of Illinois at Urbana-Champaign
210F Education Bldg., 1310 S. 6th St.
Champaign, IL 61820
217-333-8527 (phone)
217-244-7620 (fax)
jpr@illinois.edu

Abstract

In the absence of a randomized control trial, regression discontinuity (RD) designs can produce plausible estimates of the treatment effect on an outcome for individuals near a cutoff score. In the standard RD design, individuals with rating scores higher than some exogenously determined cutoff score are assigned to one treatment condition; those with rating scores below the cutoff score are assigned to an alternate treatment condition. Many education policies, however, assign treatment status on the basis of more than one rating-score dimension. We refer to this class of RD designs as "multiple rating score regression discontinuity" (MRSRD) designs. In this paper, we discuss five different approaches to estimating treatment effects using MRSRD designs (response surface RD; frontier RD; fuzzy frontier RD; distance-based RD; and binding-score RD). We discuss differences among them in terms of their estimands, applications, statistical power, and potential extensions for studying heterogeneity of treatment effects.

Regression discontinuity designs with multiple rating-score variables

Introduction

Regression discontinuity (RD) designs for inferring causality in the absence of a randomized experiment have a long history in the social sciences (see Cook, 2008) and have become increasingly popular in recent years (e.g., Cook, Shadish, & Wong, 2008; Jacob & Lefgren, 2004; *Journal of Econometrics*, 2008; Ludwig & Miller, 2007). Because the mechanism for selection into the treatment/control condition is known and observable in a regression discontinuity design, RD can provide unbiased estimates of treatment effects under much weaker assumptions than required for other quasi-experimental methods (Cook et al., 2008). Traditional RD utilizes a discontinuity in the receipt of treatment along a continuous measure (referred to as the *rating score*, *running variable*, or *forcing variable*), and estimates the treatment effect as the difference in the estimated limits of the average observed outcomes on either side of the discontinuity. Some recent examples of the application of RD designs in educational research include studies of Reading First (Gamse, Bloom, Kemple, & Jacob, 2008), Head First (Ludwig & Miller, 2007), public college admission policies (Niu & Tienda, 2009; Kane, 2003), and remedial education (Jacob & Lefgren, 2004; Matsudaira, 2008).

However, many education policies rely on more than one rating score to determine treatment status.[1] For instance, state high school exit exam policies often condition diploma receipt on student test scores in both mathematics and English language arts (e.g., Martorell, 2005; Papay, Murnane, & Willett, 2010; Reardon, Atteberry, Arshan, & Kurlaender, 2009; Ou, 2009). Similarly, rigid cutoff scores on multiple rating scales are used for determining services

[1] We note that this is distinct from one rating score variable with multiple cutoff scores resulting in multiple treatment conditions (e.g., Black, Galdo, & Smith, 2005). As long as there is one rating score—regardless of the number of cutoff scores in the rating score—we refer to this class of RDs as "single rating score RDs" or single RSRDs.

for English learners in California (Robinson, 2008, under review) and higher education financial aid programs (Kane, 2003). Likewise, school accountability policies that label schools "failing" (e.g., *No Child Left Behind*) often base the label determination on whether multiple subgroups of students each attain their annual objectives.

In such cases—where treatment assignment is determined on the basis of two (or more) continuous rating scores—the basic logic of traditional regression discontinuity applies. Nevertheless, regression discontinuity designs using multiple rating scores (hereafter, multiple rating score RD, or MRSRD) are distinctly different from RD designs using a single rating score in that the *combination* of cutoff scores attained determines treatment status. As a result, designs incorporating multiple rating-score variables raise three issues not present in the single rating score case: First, multiple rating scores may determine assignment to more than two treatment conditions. Second, rather than provide estimates of a single estimand for a single population (the effect of the treatment for individuals with rating scores near the cutoff score), MRSRD may provide estimates of multiple estimands (corresponding to the multiple possible treatment contrasts and for different subpopulations). And third, the analyst is faced with a wider range of strategies for estimating treatment effects from a multiple rating score regression discontinuity. The choice among these different strategies has important implications for precision, bias, and generalizability.

Despite considerable recent work on the statistical underpinnings and practical estimation of regression discontinuity using a single rating score (see the special issue of *Journal of Econometrics*, 2008), the current literature lacks a thorough examination of issues concerning the study of program effects when multiple cutoff scores are used to determine eligibility or

participation. In this paper, we outline these issues, and describe their implications for the estimation of treatment effects.

This paper addresses these issues and offers suggestions for implementation. In the next section, we provide a brief review of the single rating score RD estimator and then generalize the single RSRD design to the multiple rating score case. Here, we discuss how cutoffs in multiple rating score variables can create multiple treatment contrasts, leading to many possible estimands. The following section discusses several approaches to estimating average local treatment effects with multiple rating score variables. We discuss the assumptions and implementation concerns related to each approach. The next section addresses issues related to power in estimating the effects. Heterogeneity of treatment effects, which can be studied with MRSRD, is discussed in the following section. Our final section concludes with a comparative review of the MRSRD methods discussed in the paper, as well as a set of practical suggestions for analyzing data from MRSRD designs.

A brief review of the RD estimator

We frame our discussion in terms of the potential outcomes framework (see Fisher, 1935; Heckman, 1979; Holland, 1986; Neyman, 1923/1990; Rubin, 1978). First, consider the standard regression discontinuity design where treatment is assigned on the basis of a single rating score. Let R indicate the rating variable, with the cutoff score at $R = 0$, such that cases with $R \geq 0$ are assigned treatment a and cases with $R < 0$ are assigned treatment b. Each individual i has two potential outcomes, one outcome (denoted Y_i^a) that will result if the individual is assigned to

treatment a, and another (Y_i^b) that will result if he or she is assigned to treatment b.[2] The expected outcome under treatment a for individuals with $R = r$ is denoted $\bar{Y}^a(r) = E[Y|T = a; R = r]$. Under the assumption that $\bar{Y}^a(r)$ and $\bar{Y}^b(r)$ are continuous functions of r at $r = 0$, the average effect of treatment a relative to b at $r = 0$ can be written as

$$\delta(r)|(r = 0) = \lim_{r \to 0^+} \bar{Y}^a(r) - \lim_{r \to 0^-} \bar{Y}^b(r) \quad [1]$$

Under the continuity assumption, both limits on the right hand side of Equation (1) can be estimated from the observed data. The difference in the estimated limits is the regression discontinuity estimator.

In order to obtain an unbiased estimate of this difference, we need only unbiased estimates of each limit. We can obtain these from a parametric regression model, under the relatively mild assumption that we have the functional form of the model correctly specified. A general parametric version of the model is

$$Y_i = f(r_i) + \delta(T_i) + \epsilon_i, \quad [2]$$

where r is centered at the cutoff score, f is a continuous function at $r = 0$, and T is an indicator variable indicating whether $r \geq 0$ or not. We can also estimate $\delta(r)|(r = 0)$ nonparametrically, by using a smoothing estimator to obtain each of the limits in Equation (1) (Hahn, Todd, & Van der Klaauw, 2001).[3] Because the limits in Equation (1) can only be estimated from the observed

[2] Note that this formulation implicitly assumes that i's potential outcomes are independent of the treatment assignment of all other individuals, as assumption known as the "no interference between units" assumption (Cox, 1958), or the "Stable-Unit-Treatment-Value Assumption" (or SUTVA; Rubin, 1986).

[3] In practice, however, nonparametric estimators rely on linear regression models fit to data within a (narrow) bandwidth around the cutoff score, implying that even such estimators have the form shown in Equation (2) (Imbens & Lemieux, 2008).

data at $r = 0$, the regression discontinuity estimator yields an estimate of the effect of a versus b only for individuals with values of the rating score right at the cutoff score.[4]

Examples of treatment assignment on the basis of multiple rating scores

Two concrete examples will help to illustrate our discussion of the multiple rating score RD design. First, consider a case where students are assigned to remedial or summer coursework on the basis of their scores on two tests, one in mathematics and one in English language arts (ELA) (examples of such policies include the summer school program described by Jacob and Lefgren, 2004, or the assignment to remedial coursework on the basis of a student's performance on a high school exit exam in 10th grade as described by Reardon and colleagues, 2009). In this case, students who score below a given cutoff score on the math test are assigned to remedial coursework or summer school in math; students who score below a given cutoff score on the ELA test are assigned to remedial coursework in ELA. This results in four possible treatment conditions: no remedial courses; remedial math courses; remedial ELA courses; and remedial courses in both math and ELA.

Second, consider a case where students are assigned to receive special services designed for English learner (EL) students unless they demonstrate adequate mastery of both academic and conversational English, in which case they are designated as "fluent English proficient" students and receive a standard curriculum and instruction (see Robinson, 2008, under review, for a discussion of such a policy in California). In this case, students who score below the

[4] Nonetheless, the assumption that the potential outcome surfaces vary continuously with the rating score implies that the average treatment effects vary continuously with the rating score (that is, $\delta(r)$ is a continuous function at $r = 0$). If $\delta(r)$ is twice-differentiable with respect to r and $\delta''(r) \approx 0$ at the cutoff score, the estimated treatment effect at the cutoff will be approximately the same as the average effect in the region around the cutoff score. Thus, we often speak of the estimates from an RD design as generalizing to individuals with rating scores "near" the cutoff score. For precision, however, we will apply the stricter interpretation throughout this paper.

passing cutoff score on either the test of academic English or the test of conversational English receive EL services; students who score above the cutoff score on both receive standard instructional services. In this case, although there are two rating scores used to determine treatment assignment, there are only two distinct treatment conditions.

Possible estimands from multiple rating score regression discontinuity

In general, a treatment effect estimator produces an estimate of the average difference in some outcome Y that we would observe if a set of individuals in some population P were exposed to some treatment a rather than some treatment b. Defining an estimand thus requires we specify three things: (1) the outcome of interest Y, (2) the population of interest P, and (3) the treatment contrast of interest (a relative to b). In a simple randomized trial comparing two treatment conditions, we can obtain unbiased estimates of a rather general estimand (the effect of one treatment versus the other in the population of which the randomized sample is representative) or of any number of sub-population-specific estimands (the effect of one treatment versus the other in any observable sub-population of the population of which the randomized sample is representative). In a RD design with a single rating score, we can obtain unbiased estimates of only a more limited set of estimands: specifically, we can estimate the effect of one treatment versus the other in the sub-population of individuals with rating scores at (or near) the cutoff score of which the randomized sample is representative. In the case where treatment assignment is based on two or more rating variables, however, the set of possible estimands is even more constrained, as we describe below.

The logic of using regression discontinuity to estimate treatment effects when treatment is assigned on the basis of two (or more) rating variables is similar to that of the single rating

score RD. For simplicity, we will restrict our discussion and examples to the case when treatment is assigned on the basis of two rating scores, though the issues are the same with any number of rating scores. Let $R1$ and $R2$ indicate rating variables, with cutoff scores at $R1 = 0$ and $R2 = 0$. Let p denote points in the 2-dimensional space R defined by $R1$ and $R2$.

Figure 1 presents a stylized example of the joint distribution of two hypothetical rating score variables $R1$ and $R2$ used to determine treatment status. The cutoff scores on $R1$ and $R2$ are indicated by the solid lines. For generality, consider the case where there are four distinct treatment conditions, $a, b, c,$ and d. Individuals with both rating scores above their respective cutoff score (i.e., those in region A of Figure 1) are assigned to treatment a. Those with scores below the cutoff score on $R1$ but above the cutoff score on $R2$ (those in region B of Figure 1) are assigned to treatment b. Conversely, those with scores above the cutoff score on $R1$ but below the cutoff score on $R2$ (those in region D of Figure 1) are assigned to treatment d. Finally, individuals with both rating scores below their respective cutoff scores (those in region C) are assigned to treatment c. In the remediation example above, $R1$ and $R2$ are scores on the math and ELA placement tests, and treatments $a, b, c,$ and d correspond to the four possible remediation treatment conditions (respectively, no remedial courses; remedial math courses only; remedial courses in both math and ELA; and remedial ELA courses only). In the EL services example above, $R1$ and $R2$ are scores on the academic and conversational English tests; treatment a corresponds to the standard instructional program; and treatments $b, c,$ and d are identical, corresponding to the EL services treatment condition.

(Figure 1 here)

Because there are four treatment conditions, each individual has four potential outcomes, denoted $Y_i^a, Y_i^b, Y_i^c,$ and Y_i^d. We denote the expected outcome under treatment t among

individuals at p as $\bar{Y}^t(p)$. The average effect of one treatment (say a) relative to another (say b) among individuals at point p is then $\delta^{ab}(p) = \bar{Y}^a(p) - \bar{Y}^b(p)$. Then, the average effect of a relative to b in the population P is given by

$$\delta_P^{ab} = \int_{p \in R} \delta^{ab}(p)\rho(p)dp \qquad [3]$$

where $\rho(p)$ is the density of the population at point p and R is region of the two-dimensional real space containing the population P.

With a randomized experiment, we could obtain unbiased estimates of δ_P^{ab} (as well as δ_P^{ac}, δ_P^{ad}, and so on). Regression discontinuity, however, does not provide estimates that are generalizable to the full population P. Rather, RD provides unbiased estimates of the average effect of a relative to b for the subset P^{AB} of P with values of $R1$ and $R2$ that lie at the boundary between subregions A and B of R. We denote this region as $R^{A|B}$. In Figure 1, for example, a regression discontinuity estimate of the effect of a versus b would apply only to the subset of the population with values of $R1$ and $R2$ that lie along $R^{A|B}$.

More formally, two assumptions must hold in order that the multiple rating score regression discontinuity design will provide unbiased estimates of average treatment effects for the population at a treatment discontinuity. First, we assume that each average potential outcome surface $\bar{Y}^t(p)$ is a continuous function of p (that is, it is a continuous function of $R1$ and $R2$) at the boundary (or boundaries) where we are estimating treatment impacts. Second, we assume that the cutoff scores on $R1$ and $R2$ are exogenously determined; this implies that $[T_i \perp Yia, Yib, \ldots, Yid \mid p$. Let dp,q denote the Euclidean distance between points p and q in $R2$. Then, for any two treatment conditions t and u, we have the following:

$$\begin{aligned}\delta^{tu}(p) &= \bar{Y}^t(p) - \bar{Y}^u(p) \\ &= \lim_{s\to 0^+} \bar{Y}^t(q)|(d(p,q)=s) - \lim_{s\to 0^+} \bar{Y}^u(q)|(d(p,q)=s) \\ &= \lim_{s\to 0^+} \bar{Y}^t(q)|(d(p,q)=s, T=t) - \lim_{s\to 0^+} \bar{Y}^u(q)|(d(p,q)=s, T=u)\end{aligned} \quad [4]$$

As long as the observed data contain cases assigned to treatment t that are arbitrarily close to point p and cases assigned to treatment u that are arbitrarily close to point p, the limits in the last row of Equation (4) can be estimated from the observed data. Moreover, assuming the functional form of the estimator used to obtain these limits is appropriate, the estimated limits will be unbiased, yielding an unbiased estimate of the average treatment effect at p. Thus, we can obtain unbiased estimates of the average effect of treatment t versus treatment u only for points on the boundary that determines assignment to treatment conditions t and u. Generally, however, rather than estimate $\delta^{tu}(p)$ at some specific point or points along this boundary, we estimate the average value of δ^{tu} over the population at the boundary.[5]

Returning to the examples above, this implies six different potential estimands in the remedial course assignment example (refer to Figure 1):

1. The average effect of a (no remediation) versus b (math remediation) in the region $R^{A|B}$;
2. The average effect of b (math remediation) versus c (math and ELA remediation) in the region $R^{B|C}$;
3. The average effect of c (math and ELA remediation) versus d (ELA remediation) in the region $R^{C|D}$;
4. The average effect of a (no remediation) versus d (ELA remediation) in the region $R^{A|D}$;

[5] In some cases, however, we may be interested in investigating the heterogeneity of δ^{tu} across individuals with different values of p within the boundary region, a topic we return to later in this paper.

5. The average effect of a (no remediation) versus c (math and ELA remediation) at the origin;
6. The average effect of b (math remediation) versus d (ELA remediation) at the origin.

The analyses by Reardon et al. (2009) provide a useful example of several of these different estimands. They are interested in the effects of failing versus passing a high school exit exam in 10th grade. They provide four estimates, corresponding to estimands 1-4 above: the effect of failing the math exam among those with math scores at the cutoff score and ELA scores above the passing score; the effect of failing the math exam among those with math scores at the cutoff score and ELA scores below the passing score; the effect of failing the ELA exam among those with ELA scores at the cutoff score and math scores above the passing score; and the effect of failing the ELA exam among those with ELA scores at the cutoff score and math scores below the passing score.

In our second concrete example, where there were only two treatment conditions (standard instruction and EL instruction), only a single estimand is defined: the average effect of a (standard instructional services) versus b (EL instructional services) in the region defined by the union of $R^{A|B}$ and $R^{A|D}$. As above, however, we may wish to estimate the average effect of the treatment separately for the regions $R^{A|B}$ and $R^{A|D}$, particularly if we are interested in investigating the extent of heterogeneity of the effect. Kane (2003), for example, does this in his analysis of the effects of college financial aid. In his data, students receive financial aid if they meet a GPA criterion, an income criterion, and an assets criterion. In one of his analyses, inferences about the effect of the financial aid program can only be generalized to students at the GPA margin, among the subset of students who met the income and assets eligibility criteria. In

another of his analyses, inferences are generalizable to students at the income margin, given they met the asset and GPA criteria.

Estimation strategies for multiple rating score regression discontinuity

Conceptually, the analytic strategy of single or multiple rating score regression discontinuity is straightforward: we use the observed data to estimate the limits of the average potential outcomes functions at the boundary of two treatment assignment regions, and then take the difference of these estimated limits. The challenge lies in the fact that these limits must be estimated at the boundary of the observed data for each treatment condition; this requires fitting a regression model. As we discuss below, all sharp regression discontinuity estimators are based on fitting regression models of the following form to the observed data, where there are J rating scores and k distinct treatment conditions:

$$Y_i = f\left(R_i^1, R_i^2, \dots, R_i^J\right) + \sum_k \delta_k T_i^k + \mathbf{X}_i\mathbf{B} + \epsilon_i, \text{where } \left\{R_i^1, R_i^2, \dots, R_i^J\right\} \in D \subset R \quad [5]$$

Here $R_i^1, R_i^2, \dots, R_i^J$ are the J rating scores used to determine treatment status, and the T_i^k are binary variables indicating if individual i is assigned to treatment k. Within this general form, the estimators differ in two important ways: 1) the specification of the function f; and 2) the domain (D) of observations used in estimating the model.[6] The inclusion of a vector $\mathbf{X}_i$ of pre-treatment covariates in the model may increase the precision of the estimates, but is generally unnecessary, as the model is well-identified without it (Imbens & Lemieux, 2008).

[6] Even "nonparametric" regression discontinuity estimators have this form. In order to estimate the limits of the average potential outcomes functions, such estimators assume a linear average potential outcome function (so f is linear on either side of the cutoff score, though possibly with a different slope on either side) within some narrow bandwidth of the cutoff score (so D is defined as all points within some distance from the cutoff score, and the distance may differ between the two sides). In addition, they may include some kernel weighting to give more weight to observations close to the cutoff score (Imbens & Lemieux, 2008).

The choice of functional form of f used to model the average potential outcome surface may be consequential, particularly when data are relatively sparse in the region near the boundary. The ideal situation would be an ample supply of data near a boundary. In this case, we might use only those observations arbitrarily close to either side of the boundary for estimation of the limits, a strategy that minimizes the necessity of making strong assumptions about the functional form of f. Under the continuity assumption, cases on either side of a boundary will become arbitrarily similar to one another, on average, save for their treatment status, as we narrow the distance from the boundary. In the extreme, if we limit our analyses only to cases arbitrarily close to the boundary, we can analyze the data as if they were produced by a tie-breaking experiment. In the absence of very large amounts of data, however, restricting analyses to points very near to the cutoff score results in imprecise estimates of the limits, necessitating the use of data further from the cutoff score and assumptions about the functional form of the average potential outcome surfaces. Ideally, we would like estimates that are both unbiased and precise, but these goals are somewhat at odds with one another: we can gain precision by including observations further from the boundary and an assumed functional form to estimate the limits of the average potential outcomes functions, but doing so increases the potential bias in the estimated limits. We can reduce potential bias by narrowing the bandwidth, but generally at the cost of precision.

Five estimation strategies

We now introduce five possible approaches to estimating treatment effects using data from multiple rating score regression discontinuity designs. To aid understanding, we reference Figure 1 throughout our introduction to the various approaches. First, we describe *response*

surface regression discontinuity analysis. In this approach, we fit a parametric model to the full response surface, modeling treatment impacts as discontinuities in this surface at the treatment assignment boundaries. Second, we describe *frontier regression discontinuity*, in which we subset the data to estimate pairwise treatment effects by fitting single rating score regression discontinuity models to subsets of the data in adjacent subregions of R (e.g., in Figure 1, we might use only individuals in regions A and D to estimate the effect of a versus d). Third, we describe *fuzzy frontier regression discontinuity*, in which the cutoff on a single rating score serves as an instrument for treatment assignment. Fourth, we describe *distance-based regression discontinuity* which uses the distance to a point (e.g., the origin) as the rating variable for comparing available treatment contrasts at that point. Finally, we describe *binding-score regression discontinuity*. In this approach, we construct from the multiple rating scores a unidimensional rating score that perfectly predicts treatment assignment; single-rating score regression discontinuity models using this constructed score provide estimated treatment effects. Table 1 provides an overview of the estimands obtained from the various approaches, as well as a brief discussion of the approaches' advantages and disadvantages.

(Table 1 here)

1. Response surface regression discontinuity

Perhaps the most obvious way to estimate the effects of each treatment relative to the other(s) is to model the treatment effects as displacements of a multidimensional surface (see Robinson, 2008). That is, we simply fit a model of the form shown in Equation (5) above, where f is a continuous function describing the shape of the average observed potential outcomes surface and the δ coefficients indicate the average differences at the boundaries between

treatment assignment regions. Under the assumption that f has the correct functional form, this will yield unbiased estimates of the average treatment effects at the boundaries. Note the response surface RD approach simultaneously estimates each possible treatment contrast (each of the estimands described above) from a single regression model, and can be used regardless of the number of rating scores or treatment conditions.

When the multiple rating scores are used to assign individuals to one of two possible treatments ("treatment" and "control"), the typical response surface regression discontinuity model is

$$Y_i = f\left(R_i^1, R_i^2, \ldots, R_i^J\right) + \delta T_i + \epsilon_i. \quad [6]$$

In this model, f is a continuous function (typically a multidimensional polynomial surface); and T_i indicates the treatment status assigned on the basis of $\left\{R_i^1, R_i^2, \ldots, R_i^J\right\}$. Note that the model assumes that the treatment impact is unrelated to the rating scores. The estimand from this model is the average treatment impact among individuals whose vector of rating scores places them on the margin of being assigned the treatment or control condition.

When $J \geq 2$ rating scores are used to assign individuals to one of K possible treatments (where $2 \leq K \leq 2^J$), the typical response surface regression discontinuity model is

$$Y_i = f\left(R_i^1, R_i^2, \ldots, R_i^J\right) + \sum_k \delta_k T_i^k + \epsilon_i. \quad [7]$$

As above, f is a continuous function and T^k indicates assignment to treatment condition k.

The strength of the response surface regression discontinuity analysis approach is that it can use all of the available data in a relatively parsimonious way, yielding relatively precise

estimates of the treatment effect. The weakness of the method is that it requires a strong functional form assumption to estimate the treatment effect and may rely on data far from the cutoff score. If the functional form of the model is misspecified, the estimates may be biased.

2. Frontier regression discontinuity

Rather than model the multidimensional response surface, a simpler method is to subset the data by status (above or below the cutoff score) on all but one of the rating scores, and then model the discontinuity along the remaining rating score using standard single rating score RD methods. This is the approach used by Kane (2003), Papay et al. (2010), and Reardon et al. (2009). For example, to estimate the effect of treatment a versus b in Figure 1, we can limit the sample to those with $R2 \geq 0$ and use traditional single rating score regression discontinuity methods to estimate the effect of a versus b. The estimand here will be the average effect of a versus b for individuals along the boundary $R^{A|B}$.

The average local treatment effect of a relative to b (see Figure 1) for individuals who score above the cutoff score on $R2$ can be estimated by fitting the regression model

$$Y_i = f(R1_i) + \delta^{a|b} T_i^a + \epsilon_i \quad [8]$$

on the sample with $R2_i \geq 0$ and with $R1_i$ within some specified domain surrounding the cutoff score. Note that Equation (8) is a special case of Equation (5). The parameter $\delta^{a|b}$ indicates the effect of *a* relative to *b* at the frontier defined by $R^{A|B}$. Including a function of the other rating score(s) (e.g., $R2$) and covariates in the model may improve the precision of the estimates of $\delta^{a|b}$:

$$Y_i = f(R1_i) + \delta^{a|b} T_i^a + g(R2_i) + \mathbf{X}_i\mathbf{B} + \epsilon_i \quad [9]$$

We can then estimate the average effect of b versus c, c versus d, and d versus a in a similar fashion. It is important to note that these effect estimates are not necessarily comparable because each applies to a different sub-population. That is, we cannot estimate the effect of a versus c simply by adding the estimated average effects of a versus b in $R^{A|B}$ and b versus c in $R^{B|C}$ unless we assume that the effects are homogeneous across the population of interest (or at least among the populations at these two subregions). If there are only two treatment conditions, as in the EL services example given earlier, then the frontier regression discontinuity approach can provide some evidence regarding heterogeneity of effects by comparing the average effects of the treatment among those along frontier $R^{A|B}$ and those along frontier $R^{A|D}$.

One advantage of this approach is that it is straightforward. It reduces the multiple rating score regression discontinuity analysis to a set of single rating score analyses, and so relies on well-understood methods of estimating effects in single rating score regression discontinuity designs. These estimates rely on the same (relatively mild) assumptions of the single RSRD model. The frontier RD approach, however, uses only a portion of the available data for each estimate, and so may yield less precise estimates than approaches that use all available data.

3. Fuzzy frontier regression discontinuity

A modification of the frontier regression discontinuity approach is to use all of the available data to estimate the average treatment effect at a given frontier, using an instrumental variables framework. Suppose we wish to estimate the effect of treatment a versus b over the frontier $R^{A|B}$ (e.g., the effect of no remediation versus a remedial math course for students who passed the ELA test and scored at the passing margin on the math test in our remediation example above). We define an indicator variable Z_i such that $Z_i = 1$ if $R1_i \geq 0$ and $Z_i = 0$ if

$R1_i < 0$. We use Z_i as an instrument for T_i^a and estimate the equations via two-stage least squares (or some other IV estimation method):

$$T_i^a = h(R1_i) + \gamma Z_i + \boldsymbol{X}_i\boldsymbol{\Gamma} + \mu_i$$

$$Y_i = f(R1_i) + \delta^{a|b}(T_i^a) + \boldsymbol{X}_i\boldsymbol{\Delta} + \epsilon_i. \quad [10]$$

Compared to the frontier RD, this approach is has the potential to utilize more (or even all) data, because the data in regions C and D can be used to estimate the treatment effect.

However, it is important to note that this approach relies on additional assumptions. In particular, the fuzzy regression discontinuity approach requires the standard instrumental variables exclusion restriction: Z may affect Y only through T (Angrist, Imbens, & Rubin, 1996). This implies that treatments c and d be identical (or produce identical average potential outcomes at the boundary $R^{C|D}$). Formally, this requires

$$\lim_{r1\to 0^+} E[Y|R1 = r1, R2 < 0, T = c) = \lim_{r1\to 0^-} E[Y|R1 = r1, R2 < 0, T = d). \quad [11]$$

Because fuzzy regression discontinuity combines the regression discontinuity estimator and an instrumental variables estimator, its estimand is a combination of the RD and IV estimands (Hahn et al., 2001; Imbens & Lemieux, 2008; see also Trochim, 1984). In particular, fuzzy regression discontinuity yields an estimate of the local complier average treatment effect—that is, the estimated effect of treatment a versus b for those with $R1$ scores near 0 whose treatment status is affected by whether $R1$ is above or below 0.[7] Because treatment assignment to a or b is not affected by $R1$ for those with $R2 < 0$, the population to whom the estimates generalize is those on the boundary $R^{A|B}$. Thus, the fuzzy frontier regression discontinuity estimator identifies the same estimand as the frontier discontinuity estimator. It has the advantage of using more of

[7] In this IV framework, we assume no "defiers"—that is, no individuals who deliberately take the opposite treatment than the one they were assigned to. This is an innocuous assumption if the cutoff scores are strictly adhered to.

the data than the frontier regression discontinuity (because it does not discard the data with $R2 < 0$), which may increase the precision of the estimates. However, the added uncertainty inherent in instrumental variables estimates may reduce the precision of the estimates, a point we discuss below.

4. Distance-based regression discontinuity

Estimating treatment contrast *a* vs *c* (or *b* vs *d*) can be difficult with finite samples because we are comparing observations at one point with those at another point (instead of along a line, as we would if we examined treatment contrast *a* vs *b*, for example). One strategy for studying such contrasts is to construct a variable that is the Euclidean distance from point p_i to the origin. If we have two rating score variables, we can construct the new rating variable $d_i = \sqrt{R1_i^2 + R2_i^2}$.[8] Using only observations exposed to one of the two treatments of interest (e.g., only those in regions A and C), we can fit the model

$$Y_i = f(d_i) + \delta^{a|c} T_i^a + \epsilon_i \quad [12]$$

Note that we discussed the distance to the origin, but this approach could be used for estimating the effect at any point $(R1^*, R2^*, \ldots, RJ^*)$, where the distance to the point is $d_i = \sqrt{\sum_{j=1}^{J}(Rj_i - Rj^*)^2}$. However, most applications of this method will have low power for estimating the effect at any single point.

5. Binding-score regression discontinuity.

[8] This method can be extended to higher dimensions (where *J* indicates the number of dimensions) by calculating each *i*'s Euclidean distance to the origin, $d_i = \sqrt{\sum_{j=1}^{J}(Rj_i)^2}$.

When multiple rating scores determine assignment to only two treatment conditions, one can construct a new rating score that alone perfectly determines treatment assignment. For example, in our earlier example regarding reclassification of English Learners, students receive one treatment (they are reclassified) if they score at or above a given cutoff score on each of 5 separate tests; otherwise they receive the control condition (not reclassified). No one of the 5 scores alone perfectly determines treatment assignment, but we can construct a new rating variable, M_i, defined as the minimum of the 5 test scores (where each score is first centered around its cutoff score), that does perfectly determine assignment:

$$M_i = \min(R1_i, R2_i, \ldots, R5_i). \quad [13]$$

M is a continuous, observable variable, defined so that $T_i = 1$ if $M_i \geq 0$ and $T_i = 0$ if $M_i < 0$. Given M, we can use single rating score regression discontinuity methods to estimate the effect of the treatment for those values of $M \approx 0$ (those whose lowest score among the five was right at the margin of passing). That is, we then fit a model of the form

$$Y_i = f(M_i) + \delta T_i + \epsilon_i \quad [14]$$

to estimate the effect of the treatment.

More generally, we can use binding-score regression discontinuity by constructing a new rating score $M_i = g(R1_i, \ldots, RJ_i)$ such that $g(0, \ldots, 0) = 0$ and g is monotonic in each Rj. This raises the question of what function to choose for g. One strategy has been to construct Mas the minimum of the standardized rating scores, based on the rationale that this makes the distance from the cutoff score comparable for each of the rating scores (Martorell, 2005; Robinson, under review). Although any function g defined as above will provide an unbiased estimate of the treatment effect (assuming we can identify a correct functional form for the model relating Y to M), the choice of g may affect the power of the regression discontinuity model (because it will

affect the strength of the correlation between M and Y, conditional on T, which affects the power of the regression discontinuity estimator (Bloom, 2009). There is as yet little empirical or theoretical guidance, however, on how to pick a function f that will maximize the precision of the binding score design.

It may be useful to include the individual rating scores in the model as covariates to increase the precision of the estimates. In this case, the model becomes

$$\begin{aligned} Y_i &= f(M_i) + h(R1_i, R2_i, \ldots, RJ_i) + \delta T_i + \epsilon_i \\ &= f\big(g(R1_i, R2_i, \ldots, RJ_i)\big) + h(R1_i, R2, \ldots, RJ_i) + \delta T_i + \epsilon_i \\ &= f'(R1_i, R2_i, \ldots, RJ_i) + \delta T_i + \epsilon_i. \end{aligned} \quad [15]$$

Thus, the binding score regression discontinuity model is a special case of the response surface regression discontinuity model, where the function describing the response surface includes a minimizing or maximizing function g.

The binding-score regression discontinuity model is appealing because it may use all of the available data to estimate the treatment effect, but does so by parsimoniously reframing the multidimensional vector of rating scores into a single dimension that alone determines treatment status.

Statistical power of MRSRD designs

The goal with any RD design is to obtain a *precise* unbiased estimate of the treatment effect (Imbens & Lemieux, 2008; Schochet, 2009); this is also true for MRSRD. The standard error of the treatment effect estimate [$se\big(\hat{\delta}^{RD}\big)$] in any RD design with a single treatment contrast is a function of sample size (N), the proportion of the sample assigned to the treatment

(P), the within-group sample variance in the outcome variable ($\hat{\sigma}^2_{Y|T}$), the proportion of variance in treatment status explained by the rating-score variable(s) and any covariates ($R^2_{T.\mathbf{R},\mathbf{X}}$), and the proportion of error variance left unexplained by the rating score(s) and covariates ($1 - R^2_{Y.\mathbf{R},\mathbf{X}|T}$) (see Bloom, 2009; Schochet, 2009, for recent detailed discussions):

$$se\left(\hat{\delta}^{RD}\right) = \sqrt{\frac{\left(\hat{\sigma}^2_{Y|T}\right)\left(1 - R^2_{Y.\mathbf{R},\mathbf{X}|T}\right)}{NP(1-P)\left(1 - R^2_{T.\mathbf{R},\mathbf{X}}\right)}} \quad [16]$$

When there is a single treatment contrast (that is, when multiple rating scores determine assignment to one of only two treatments, as in the reclassification example above), both the surface RD and the binding-score RD methods estimate the same estimand—the average treatment effect for those whose combination of rating scores places them at the boundary of the two treatment assignment regions. Assuming both methods use the same subsample of the available data, N and P will be the same in both cases. The relative power of the two methods will therefore depend on the two R^2 terms. The relative magnitude of the $R^2_{Y.\mathbf{R},\mathbf{X}|T}$ term will depend on whether the surface function f in Equation (6) fits the data better than the binding score function f' in Equation (15).[9] The relative magnitude of the $R^2_{T.\mathbf{R},\mathbf{X}}$ term will generally be smaller in the surface RD method, all else equal, because the binding score alone captures more of the variance in treatment status than can the multiple rating scores alone (absent some interaction or higher-order terms of the rating scores).[10] Because a higher value of $R^2_{T.\mathbf{R},\mathbf{X}}$ implies

[9] Different functions may yield the same unbiased effect estimates, because for unbiasedness, we require only that the limits of the functions approach the same values at the treatment assignment boundary. The functional forms may predict different values elsewhere in the region R. As a result, different functional forms may yield unbiased estimates but have different values of $R^2_{Y.\mathbf{R},\mathbf{X}|T}$.

[10] To see this, note that if treatment is defined such that $T_i = 1\{M_i \geq 0\}$, where $M = \min\{R1, R2\}$, and we regress $T_i = \beta_0 + \beta_1 M_i + \beta_2 R1_i + \beta_3 R2_i + e_i$, the expected values of $\hat{\beta}_2$ and $\hat{\beta}_3$ are 0 (because the rating scores tell us nothing about treatment assignment once we know M). However, the expected value of $\hat{\beta}_1$ will not be zero. Thus,

a larger standard error, the surface RD design will generally yield more precise treatment effect estimates than the binding score RD design, assuming both models use the same subsample of data and fit the data equally well (i.e., have the same residual variance).

In addition, when there is a single treatment contrast, both the frontier RD and the fuzzy frontier RD methods can be used to estimate the effect of the treatment for the subset of the population who score above the cutoff on all but one rating score and who score at or near the cutoff on the remaining rating score. In this case the relative precision of the frontier and fuzzy frontier methods depends on multiple factors. In particular, the standard error of the fuzzy RD estimate will equal the standard error of the intent-to-treat (ITT) estimate divided by the estimate of γ from the first-stage equation in model (10).[11] Because $\hat{\gamma} \leq 1$, the standard error of the fuzzy frontier RD treatment effect estimate will never be smaller than that of the ITT estimate.

The standard error of the intent-to-treat estimate from the fuzzy frontier method, however, may be larger or smaller than that of the frontier RD method. To see this, note that the standard error of the ITT estimate is given by Equation (16) (replacing T by Z in the equation). This standard error may differ from the standard error of the frontier RD estimate for several reasons. First, the fuzzy frontier RD method uses a larger subsample of the data, which will tend to reduce the standard error of the relative to the sharp frontier method, all else equal. Second, the value of P may differ between the two methods, because the subsample of data used in the estimation may differ. Because the standard error is minimized, all else equal, when $P = 0.5$, the inclusion of additional observations may tend to decrease or increase precision, depending on

---

the expected R^2 from this model will be equal to that of a model that includes only M as a predictor of treatment status, but greater than one that includes only $R1$ and $R2$ as (linear) predictors of treatment status.

[11] The intent-to-treat estimate is the coefficient $\hat{\beta}$ from the model $Y_i = g(R1_i) + \beta Z_i + \boldsymbol{X}_i\mathbf{B} + e_i$. The Wald IV estimator of the effect of T on Y is $\hat{\delta} = \hat{\beta}/\hat{\gamma}$ (where $\hat{\gamma}$ is estimated from the first stage of Equation (10) and the standard error of the estimate is $se(\hat{\delta}) = se(\hat{\beta})/\hat{\gamma}$. The coefficient $\hat{\gamma}$ is interpreted as the proportion of observation with $R1 \approx 0$ who have $R1 \geq 0$—it is the proportion of those individuals at the cutoff score of $R1$ whose treatment status is determined by $R1$.

whether the added observations increase or decrease the balance of the treatment and control group sizes. Third, the two R^2 terms may differ between the two models. Depending on the particulars of the joint distribution of $R1$ and $R2$ (and $\mathbf{X}$), $R^2_{T.\mathbf{R},\mathbf{X}}$ may be larger or smaller than $R^2_{Z.\mathbf{R},\mathbf{X}}$. As a result, the fuzzy frontier ITT estimates may be more or less precise than the frontier RD estimates, implying that the fuzzy frontier RD effect estimates may likewise be either more or less precise than the frontier estimates.

It may seem counterintuitive that the fuzzy RD estimates can be made more precise than the frontier RD estimates. The inclusion of additional observations in which treatment assignment does not vary would seem to add no additional information regarding the treatment effect. However, the additional observations may substantially improve the precision of the estimated limits of the regression function as it approaches the cutoff score. For example, suppose we are interested in estimating $\delta^{a|b}$ (see Figure 1), and suppose $R2$ was uncorrelated with Y, conditional on $R1$. Then the inclusion of individuals from region C would improve our ability to estimate the average value of Y for those with $R1 = 0$ and $Z = 0$, because those in C would have identical values of Y, in expectation, to those in region B, conditional on $R1$. Likewise, the inclusion of individuals from region D would improve our ability to estimate the average value of Y for those with $R1 = 0$ and $Z = 1$. If the increased precision gained by increasing the sample size this way were greater than the reduction in precision due to the division by $\hat{\gamma}$, the fuzzy RD estimates will be more precise than the frontier RD estimates.

Checking the assumptions MRSRD models

With both single and multiple rating score RD designs, we assume that 1) the cutoff score(s) determining treatment assignment is/are exogenously set; 2) potential outcomes are

continuous functions of the rating score(s) at the cutoff score(s); and 3) the functional form of the model is correctly specified. Imbens and Lemieux (2008) provide a detailed description of assumption checking in the single rating score RD context. We build upon their discussion to extend these checks to the case of MRSRD. In particular, we suggest assessing the assumptions as they apply to each separate frontier at which there is a treatment assignment discontinuity, regardless of which of the analytic methods are used to estimate the treatment effects. This enables the researcher to verify the plausibility of the assumptions at each treatment assignment threshold.

In order to assess the plausibility of the assumption that the cutoff scores are exogenously determined, McCrary (2008) suggests (for the single rating score RD) checking that the density of observations be similar on either side of the cutoff score. In the MRSRD design, it is useful to compare the density of observations near each of the cutoff scores. For example, if there are two rating scores and four distinct treatment conditions (as shown in Figure 1), this entails four separate comparisons (one comparing the density of observations on either side of frontier $R^{A|B}$, one comparing the density on either side of frontier $R^{B|C}$, and so on). In general, if there are J rating scores and 2^J treatment conditions, this will entail $J \cdot 2^{J-1}$ separate comparisons. Evidence of discontinuity in the density of observations across any of the frontiers suggests the possibility of manipulation of the rating scores and thus, a failure of the exogeneity assumption.

To assess the plausibility of the second assumption—continuity of potential outcomes—we may examine the continuity of observable (pre-treatment) variables presumed to be related to the potential outcomes. In practice, this is done by checking that each such variable exhibits no discontinuity at each of the $J \cdot 2^{J-1}$ frontiers. We check this by fitting a set of frontier regression discontinuity models, each with a covariate X as the dependent variable, and testing the null

hypothesis that $\lim_{r\to 0^+} E[X|R=r] = \lim_{r\to 0^-} E[X|R=r]$ for each X at each frontier. It may be useful to include the other rating scores in the set of covariates **X**, as they should not vary sharply at any frontier.[12]

The third assumption is that the functional form of the model is correctly specified. As Imbens and Lemieux (2008) point out, this is best done through visual inspection of a plot of the relationship between the outcome variable and the rating score. For each of the methods that reduce the multidimensional rating score space to a single rating score dimension (i.e., each of the methods except for the response surface RD method), this can be done straightforwardly with the methods described by Imbens and Lemieux (2008). When using the response surface RD approach, however, identifying the correct functional form in a multidimensional space through graphical analysis can be difficult. In this case, we propose a visual inspection technique for assessing the appropriateness of the functional form of the surface RD model.

Note that if the functional form of the model is correct, the residuals should have an expected value of 0 at each point in the multidimensional rating score space. We check this as follows. First, given a fitted surface RD model, predict the residuals. For each frontier where there is a treatment condition discontinuity, (e.g., $R^{A|B}$ in Figure 1), consider the sample of observations that are in the adjacent regions of the rating score space (e.g., regions A and B in Figure 1). Next, using this sample, construct a plot of the residuals from the surface RD model against the single rating score that determines treatment assignment at that frontier (e.g., $R1$ if we are examining the frontier at $R^{A|B}$). If the functional form of the model is correct, the residuals should have a mean value of 0 at each value of the rating score. Most importantly, the residuals should have a mean value of 0 as they approach the cutoff score from both the left and

[12] i.e., at the frontier defined by $R1 = 0; R2 > 0$ (frontier $R^{A|B}$ in Figure 1), it should be true that $\lim_{R1\to 0^-}(\overline{R2}) = \lim_{R1\to 0^+}(\overline{R2})$.

the right. This can be checked by fitting separate nonparametric smoothed curves through the data on either side of the cutoff score or by plotting the mean value of the residual within small bins across the range of the rating score. We construct these plots for each of the frontiers in order to assess the model fit over the entire region of the sample. Deviation of the mean value of the residual from 0—particularly at the cutoff score—suggests the response surface RD model is not fit correctly in this region of the data. Ideally, there should be no noticeable deviation from average residuals of zero throughout the domain, though the most crucial region for no deviation is around any cutoff score at which there is a treatment discontinuity.

Heterogeneity of average local treatment effects

The MRSRD designs described above estimate local average treatment effects in the region near the cutoff score of one or more rating scores. These methods can be extended to investigate the extent to which treatment effects vary in relation to another rating score. As with the study of average treatment effects, the researcher has the option of a range of approaches, each estimating a different estimand. We present stylized examples of these options in the case of the surface RD and the frontier RD methods; the extension to the other methods is similar.

1. Response surface regression discontinuity

With the response surface RD method, we can study how the relationship between the outcome and each rating-score variable changes at different values of the rating score (i.e., $\partial\delta/\partial Rj$), for all js simultaneously. For simplicity, we restrict our discussion to the case of only one treatment and one control condition, but the same principles apply to cases of additional conditions. Beginning with Equation (6), each Rj is interacted with T, yielding:

$$Y_i = f\left(R_i^1, R_i^2, \ldots, R_i^J\right) + \delta T_i + \sum_{j=1}^{J} \gamma_j (T_i \cdot Rj_i) + \epsilon_i. \quad [17]$$

From Equation (17), δ is now the treatment effect at the intersection of the cutoff scores, rather than the average treatment effect. This coefficient is likely to be uninteresting, unless one is interested in, for example, the effect of a policy on the type of student who barely meets all of the cutoff scores. Of greater interest are the coefficients $\gamma_1, \gamma_2, \ldots \gamma_J$, which represent the relationships between outcomes and scoring higher (or lower) on one cutoff score while just attaining the other cutoff scores.

2. Frontier regression discontinuity

Although heterogeneity can theoretically be estimated along all rating-score variables with the response surface RD, the frontier RD requires subsetting the data by subpopulations and estimating one heterogeneity estimate at a time. We can investigate heterogeneity of effects within a given subpopulation, such as individuals who attained the cutoff score on $R2$ (in the case of two rating score variables, $R1$ and $R2$)—those in regions A and B of Figure 1. To do this, we estimate models such as:

$$Y_i = f(R1_i, R2_i) + \delta T_i + \gamma (T_i \cdot R2_i) + \epsilon_i \quad [18]$$

where *T* indicates $R1 \geq 0$. Now δ indicates the average effect of *a* versus *b* for individuals with $R1 \approx 0$; and γ indicates the extent to which the average effect of *a* versus *b* varies linearly with $R2$ (within the subpopulation of individuals near $R^{A|B}$).

Conclusions

Multiple rating-score regression discontinuity (MRSRD) designs are common in education. In principle, they provide an opportunity to obtain unbiased estimates of treatment effects, using the same logic as single rating score RD designs. However, MRSRD designs contain some added complexity not present in single RD designs. In particular, they present the researcher with multiple possible estimands and multiple estimation strategies.

In this paper, we have presented five approaches to estimating treatment effects in the multiple rating score RD design: response surface RD, frontier RD, fuzzy frontier RD, distance-based RD, and binding-score RD. These approaches differ in their estimands and their statistical power; they differ also in the extent to which the assumptions on which they rely are easily assessed. As we have noted, each approach has advantages as well as shortcomings.

In determining which analytic approach(es) to use in MRSRD design analyses, the first thing one should consider is the estimand of interest. Depending on how many distinct treatment conditions are determined by the multiple rating scores, and depending on what subpopulation is of most interest, the researcher may desire to estimate different parameters. Depending on the choice of estimand, one or more different MRSRD strategies may be useful. The choice among these strategies should be driven by considerations of statistical power and the ease with which the identifying assumptions of the RD design are assessed.

Regardless of the strategy used, researchers should assess the extent to which the data appear to support the assumption of exogenous determination of the rating scores and cutoff scores, the assumption that potential outcomes are continuous as the cutoff scores, and the assumption that the functional form of the model is correct. Each of these assumptions are most easily assessed using frontier RD methods, because subsetting the data by frontier reduces the assumption checking process to a series of well-defined single rating score RD designs, where

methods of checking the assumptions are well-defined (see Imbens & Lemieux 2008). However, if a response surface RD model is used to estimate the treatment effects, checking the functional form assumptions can be done by inspecting the residuals from the fitted model, as we describe above.

Finally, an important feature of MRSRD is the ability to study heterogeneity of treatment effects. This feature allows us to simultaneously compare treatment effects along multiple cutoff scores/dimensions, and to potentially identify situations where policies and instructional practices are more effective. Just as different MRSRD approaches can yield different estimands in the study of local average treatment effects, they can address different questions regarding heterogeneity. Both the frontier RD and response surface RD can be used to study how the slope of the effect of the treatment received for attaining one cutoff score differs along another rating score dimension.

Although this paper describes a set of strategies for analyzing data from MRSRD designs, several issues in the analysis of such designs deserve further attention. First, the development of algorithms for the application of non-parametric methods to response surface RD models would be useful. Recent developments in the use of non-parametric models (Hahn et al., 2001) and methods of optimal bandwidth selection (Ludwig & Miller, 2007; Imbens & Lemieux, 2008) may be extended to fit multidimensional surface RD models, as suggested by Papay, Murnane, and Willett (2009). Second, although our discussion here has focused on the analysis of data from naturally-occurring MRSRD studies, the lessons from this may be used to develop guidelines for designing MRSRD studies to optimize precision. Given the joint distribution of the rating scores and the correlations between the rating scores and outcome variable, choices regarding the number of rating scores, the location of cutoff scores, and sample size will affect

the precision of impact estimates from an MRSRD design. As such designs become more common in education research, a better understanding of their effects will aid in designing appropriately-powered studies.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. R. (1996). Identification of causal effects using instrumental variables." *Journal of the American Statistical Association, 91*(434), 444-472.

Black, D., Galdo, J., & Smith, J. (2005). *Estimating the selection bias of the regression discontinuity design using a tie-breaking experiment*. Working Paper. Syracuse, NY: Department of Economics and Center for Policy Research, Syracuse University.

Bloom. (2009). *Modern regression discontinuity analysis.* New York, NY: MDRC.

Cook, T. D. (2008). "Waiting for life to arrive": A history of regression-discontinuity design in Psychology, Statistics, and Economics. *Journal of Econometrics, 142*(2), 636-654.

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

Cox, D. R. (1958). *The planning of experiments*. New York, NY: Wiley.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*(1), 39-82.

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T., (2008). *Reading First Impact Study: Interim Report* (NCEE 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Heckman, J. J.. (1979). Sample selection bias as a misspecification error. *Econometrica, 47,* 153-161.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-968.

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

Kane, T. J. (2003). *A quasi-experimental estimate of the impact of financial aid on college-going.* NBER working paper no. 9703. Retrieved December 14, 2008, from http://www.nber.org/papers/w9703.

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

Journal of Econometrics (2008), *142*(2). (Special issue on regression discontinuity).

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

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics, 142*(2), 829-850.

Martorell, F. (2005). *Do high school graduation exams matter? Evaluating the effects of exit exam performance on student outcomes.* Berkeley, CA: Unpublished working paper.

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

Neyman, J. S. (1923/1990). On the application of probability theory to agricultural experiments. *Statistical Science, 5*(4), 465-480.

Niu, S. X., & Tienda, M. (2009). The impact of the Texas top ten percent law on college enrollment: A regression discontinuity approach. *Journal of Policy Analysis and Management, 29*(1), 84-110.

Ou, D. (2009). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review, 29*(2), 171-186.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2009). The price of just failing: Consequences of high school exit examinations for urban students in Massachusetts. Paper presented at the annual conference of the Society for Research on Educational Effectiveness, Crystal City, VA, March 3, 2009.

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*(1), 5-23.

Reardon, S. F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009). *Effects of the California high school exit exam on student persistence, achievement, and graduation*. IREPP Working paper 2009-12. Stanford, CA: Institute for Research on Education Policy and Practice, Stanford University.

Robinson, J. P. (2008). *Essays on the effectiveness of policies and practices for reducing cognitive gaps between linguistic groups and socioeconomic groups.* Stanford University doctoral dissertation.

Robinson, J. P. (under review). *Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables.*

Rubin, D. B. (1978). Bayesian inference and causal effects: The role of randomization. *The Annals of Statistics, 6*(1), 34-58.

Rubin, D. B. (1986). Which ifs have causal answers? Comment on "Statistics and causal inference" by P. W. Holland. *Journal of the American Statistical Association, 81*, 961-962.

Schochet, P. Z. (2009). Statistical Power for Regression Discontinuity Designs in Education Evaluations. *Journal of Educational and Behavioral Statistics, 34*(2), 238-266.

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

Trochim, W. (1984). *Research design for program evaluation: The regression-discontinuity design.* Beverly Hills, CA: Sage Publications.

Table 1. Potential analytic scenarios and suggested methods for an MRSRD design

Number of treatment conditions	**Method of analysis**	**Estimand(s)**	**Example (on Figure 1)**	**Advantages**	**Disadvantages**
2 (e.g., a single "treatment" and single "control" condition, such as in the reclassification example of "reclassified" or "not reclassified")	Response surface RD	Average treatment effect among the population at frontiers of the treatment assignment region	Average effect among those at thresholds $R^{A\|B}$ and $R^{A\|D}$	Potentially most precise method, because it can use all available data.	Does not reduce to the familiar case of a single RSRD; modeling can be complex, and assumption checks can grow in number as dimensionality of the RSs grows; must be careful to not restrict data to unusual or artificial regions
	Binding-score RD	Average treatment effect among the population at frontiers of the treatment assignment region	Average effect among those at thresholds $R^{A\|B}$ and $R^{A\|D}$	Reduces the dimensionality to a single RSRD; uses all data	May have less power than surface RD; choice of scaling of rating scores may affect estimates
	Frontier RD	Local average treatment effect at a single frontier	Average effect along $R^{A\|B}$ (estimated using only data from regions A and B)	Reduces the dimensionality to a single RSRD; readily interpretable estimand	May not maximize power, because uses data from only two regions
	Fuzzy frontier RD	Local average treatment effect at a single frontier	Average effect along $R^{A\|B}$ (estimated using	May have greater precision than frontier RD under	May have lower precision than frontier RD;

			data from regions A, B, C, and D)	some circumstances	modeling may be more complex than frontier RD; requires that *c* and *d* are identical treatments
3 or more (as in the remediation example)	Response surface RD	Average treatment effects at each frontier that determines a treatment assignment discontinuity	Effects at $R^{A\|B}$, $R^{A\|D}$, $R^{B\|C}$, and $R^{C\|D}$	All treatment effects can be estimated in a single model, maximizing power	Model can become very complex as number of rating scores and treatment conditions grows; functional form validity hard to verify
	Frontier RD	Local average treatment effect at a single frontier	Average effect along $R^{A\|B}$ (estimated using only data from regions A and B)	Reduces the dimensionality to a single RSRD; readily interpretable estimand	Must estimate each treatment effect separately; may not maximize power, because uses data from only two regions
	Fuzzy frontier RD	Local average treatment effect at a single frontier	Average effect along $R^{A\|B}$ (estimated using data from regions A, B, C, and D)	May have greater precision than frontier RD under some circumstances	Not recommended. Requires that *c* and *d* are identical treatments (or produce identical potential outcomes); otherwise estimates will be biased
	Distance-based RD	Average treatment effect at the origin	Effect at the origin (estimated using	Only method of obtaining effect of	Low power (unless very high density of

	data from regions A and C)	treatment *a* versus *c*	observations near the origin); estimand applies to very limited population

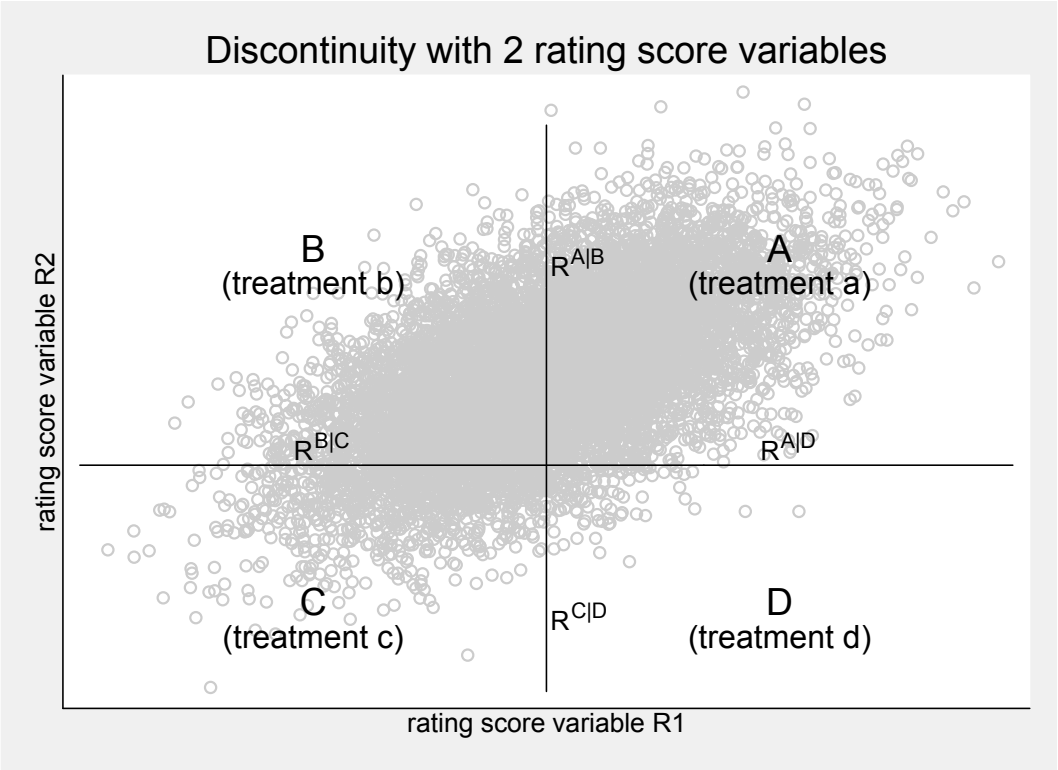

Figure 1. Discontinuity with 2 rating score variables.