Estimating Causal Effects of Education Interventions Using a Two-Rating Regression Discontinuity Design

Lessons from a Simulation Study

Kristin E. Porter
MDRC

Sean F. Reardon
Stanford University

Fatih Unlu
Abt Associates

Howard S. Bloom
MDRC

Joseph P. Robinson-Cimpian
University of Illinois at Urbana-Champaign

This working paper was supported by Grant R305D100027 to MDRC
from the Institute of Education Sciences, U.S. Department of Education.

November 2014

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

Acknowledgments

This paper was supported by Grant Number R305D100027 from the Institute of Education Sciences of the U.S. Department of Education. The authors would like to thank Asya Magazinnik, Andrea Shane, Sara Staszak, and Rebecca Coven for their help with programming, tables and figures, and report production. All shortcomings herein are, of course, solely the authors' responsibility.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, The Harry and Jeanette Weinberg Foundation, Inc., The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Abstract

A valuable extension of the single-rating regression discontinuity design (RDD) is a multiple-rating RDD (MRRDD). To date, four main methods have been used to estimate average treatment effects at the multiple treatment frontiers of an MRRDD: the “surface” method, the “frontier” method, the “binding-score” method, and the “fuzzy instrumental variables” method. This paper uses a series of simulations to evaluate the relative performance of each of these four methods under a variety of different data-generating models. Focusing on a two-rating RDD (2RRDD), we compare the methods in terms of their bias, precision, and mean squared error when implemented as they most likely would be in practice — using optimal bandwidth selection. We also apply the lessons learned from the simulations to a real-world example that uses data from a study of an English learner reclassification policy. Overall, this paper makes valuable contributions to the literature on MRRDDs in that it makes concrete recommendations for choosing among MRRDD estimation methods, for implementing any chosen method using local linear regression, and for providing accurate statistical inferences.

Contents

Section

Appendix

List of Exhibits

Table

Figure

1. Introduction: The Opportunities and Challenges of Multiple-Rating Regression Discontinuity Designs

In recent years, the regression discontinuity design (RDD) has gained widespread recognition as a quasi-experimental method that, when used correctly, can produce internally valid estimates of causal effects of a treatment, a program, or an intervention (Cook, Shadish, and Wong, 2008; Imbens and Lemieux, 2008b; Jacob and Lefgren, 2004; Ludwig and Miller, 2007). In a traditional RDD, subjects (e.g., students or schools) are rated according to a numeric index (e.g., a test score, composite performance indicator, or poverty measure) and treatment assignment is determined by whether one's rating falls above or below an exogenously defined cut-point value of the rating. The parameter of interest in a traditional RDD is the average treatment effect *at the cut-point*, or "treatment frontier," which represents the average effect for the subpopulation of individuals whose ratings equal the cut-point value. Research on the statistical properties of the RDD has provided theoretical justification and empirical verification of its internal validity (Hahn, Todd, and Van Der Klaauw, 2001; Imbens and Lemieux, 2008a; Lee, 2008; Rubin, 1977). This validity derives from the fact that an RDD can be thought of as relying on the equivalent of a randomized controlled trial (RCT) at its cut-point (Goldberger, 1972a, 1972b; Mosteller, 1990).

RDDs have been used to estimate causal effects in a variety of contexts; for a list of more than 75 studies in the contexts of education, labor markets, political economy, health, crime, and more, see Lee and Lemieux (2010). Specifically in education research, recent studies that have used a traditional RDD include those analyzing the effects of Reading First (Gamse, Bloom, Kemple, and Jacob, 2008), Head Start (Ludwig and Miller, 2007), public college admission policies (Kane, 2003; Niu and Tienda, 2009), and remedial education (Jacob and Lefgren, 2004; Matsudaira, 2008).

In many interventions, however, assignment is determined by more than one rating. Therefore, a valuable extension of the single-rating RDD is a *multiple-rating* RDD (MRRDD). For example, a *two-rating* RDD provides the opportunity to study the effects of an education intervention that is assigned to students based on both their reading *and* math scores. To date, MRRDDs have been used to study the effects of state high school exit exam policies (e.g., Martorell, 2005; Ou, 2009; Papay, Murnane, and Willett, 2010; Reardon, Arshan, Atteberry, and Kurlaender, 2010), as well as the effects of services for English learners (Robinson, 2008, 2011) and of higher education financial aid offers (Kane, 2003).

While a traditional RDD targets a single average treatment effect at a single treatment frontier, an MRRDD has multiple treatment frontiers, which makes it possible to estimate multiple average treatment effects (one for each frontier). These different effects are defined for different treatment contrasts and different subpopulations. When estimating these effects, analysts face the same well-documented challenges associated with using a single-rating RDD (Imbens and Lemieux, 2008b), plus the additional challenge of deciding among alternative MRRDD estimation methods.

To date, four main methods have been used to estimate average treatment effects at the multiple treatment frontiers of an MRRDD: the "surface" method, the "frontier" method, the "binding-score" method, and the "fuzzy instrumental variables (IV)" method.[1] However, the theoretical properties of these methods are not well understood, which has presented a challenge for applied researchers (Gill, Lockwood, Martorell, Setodji, and Booker, 2008; Jacob and Lefgren, 2004; Kane, 2003; Martorell, 2005; Robinson, 2008; Weinbaum and Weiss, 2009).

To help researchers better understand the trade-offs among these methods, several recent papers have begun to explore their statistical properties (Papay, Willett, and Murnane, 2011; Reardon and Robinson, 2012; Wong, Steiner, and Cook, 2013). These papers have made significant contributions by demonstrating, among other things, that all four methods can produce unbiased impact estimators under the right conditions. However, much further investigation is needed to help researchers make informed choices about which MRRDD estimation methods to use for a given data set. Guidance is also needed with respect to implementing each estimation method in practice.

The goal of this paper is to provide such assistance. To minimize the complexity of our presentation, we focus on the simplest type of MRRDD — a *two-rating* RDD (2RRDD) with a single treatment condition. For this design, we use a series of simulations to evaluate the performance of each of the four MRRDD estimation methods under a variety of different data-generating models. Specifically, we evaluate the methods in terms of their relative bias, precision, and mean squared error when implemented as they most likely would be in practice — using local linear regression, which uses data within an "optimal" bandwidth around the cut-point that defines each frontier.[2] In practice, the true functional form of the relationship between the outcome and ratings is unknown. Therefore, we consider the modifications needed to adopt

[1] Reardon and Robinson (2012) discuss a fifth MRRDD estimation method, which they refer to as the "distance-based" method. However, because it targets a different effect — the average treatment effect at the intersection of the two cut-points, or the origin — we do not include the method in our investigations.

[2] We also examine each method's relative precision when implemented under conditions that are theoretically ideal but unlikely to exist in practice — specifically, when the functional forms of the data-generating models are known. We show that with correct model specification, all of the estimators are unbiased, although their precision varies substantially due to differences in how much information about the assignment mechanism and the subsample of the data they use and the complexity of their functional forms. The results of these analyses are summarized later and presented in Appendix B.

the strategy of optimal bandwidth selection that is now used widely for single-rating RDDs (Imbens and Lemieux, 2008a) for use with each MRRDD method.

In addition, we assess the accuracy of statistical inferences when standard methods are used for making these inferences for each of these methods. Optimal bandwidth selection adds uncertainty to the MRRDD estimator for each method, which is typically ignored when making statistical inferences.[3] That is, the sampling variance of the MRRDD treatment effect estimators is generally estimated with standard ordinary least squares (OLS) or IV methods. We explore this issue using simulations and examine the effect of accounting for the added uncertainty produced on the estimated standard errors. This contribution applies to single-rating RDDs as well.

We supplement these simulation analyses with an analysis of a 2RRDD that relies on an existing education data set. This analysis shows how to apply the lessons learned from the simulation study to the analysis of an actual data set.

The remainder of this paper is organized as follows. Section 2 introduces the 2RRDD that is the focus of our analysis, describes the parameters to be estimated for this design, and lists the conditions necessary for identifying those parameters. Section 3 presents an overview of the four MRRDD methods as they apply to a 2RRDD design. Section 4 describes the process used to generate the data sets that are used to assess the four methods. Section 5 describes how the four methods were implemented as an analyst would proceed in practice, by estimating a local linear regression from data for an optimal bandwidth around each cut-point. Section 6 presents the results of our simulations. Section 7 presents the application of lessons learned from the simulation analyses to an existing education data set. Section 8 summarizes the main lessons we learned about a 2RRDD and considers how these lessons generalize to other MRRDDs.

[3] For example, see Papay, Murnane, and Willett (2010) and Reardon, Arshan, Atteberry, and Kurlaender (2010). We further note that in their guide to practice on RDDs, Imbens and Lemieux (2008a) state that standard least squares methods can be used in conjunction with local linear regression. The What Works Clearinghouse standards for RDDs do not mention the need to incorporate the uncertainty added by bandwidth selection into standard error estimation (Schochet et al., 2010).

2. Research Design, Parameters of Interest, and Conditions for Identification

2.1 Research Design

As noted, this paper focuses on a two-rating regression discontinuity design (2RRDD), for which subjects (e.g., students) are assigned to an intervention based on their values of two ratings, such as a reading test score and a math test score. Our design of interest, illustrated by Figure 1, has subjects assigned to a *single* treatment if they fall below an exogenous cut-point on *one or both* of the two ratings. For example, students may be assigned to an after-school program if they score below an exogenous cut-point on a math test and/or if they score below an exogenous cut-point on a reading test.

Let $R1 =$ rating 1 and $R2 =$ rating 2, and let $c1 =$ an exogenous cut-point for rating 1 and $c2 =$ an exogenous cut-point for rating 2. The treatment assignment indicator, T, is equal to 1 if $R1 < c1$ and/or $R2 < c2$. Therefore, subjects with a combination of ratings that fall into the upper right quadrant of Figure 1 are in the comparison group, while subjects with a combination of ratings that fall into any of the other three quadrants are in the treatment group.[4]

Other 2RRDDs may include two or three different treatment conditions. For example, students who fail a math test are assigned to an after-school program in math, students who fail a reading test are assigned to an after-school program in reading, and students who fail both tests are assigned to both after-school programs. Such designs are not considered in this paper; for a discussion of them, see Reardon and Robinson (2012).

2.2 Parameters of Interest

In a single-rating RDD, the main parameter of interest is the average treatment effect at the cut-point value (c) of the single rating (R) or, equivalently, the average treatment effect at the frontier that separates treatment and comparison group members, individuals who just barely miss or make the threshold for treatment assignment. For this subgroup, the difference in average outcomes among those immediately on either side of the cut-point can provide an unbiased estimator of the average treatment effect.

[4] This design is the flip side of one in which subjects are assigned to an intervention if and only if they fall below a cut-point on *both* of two ratings (i.e., the treatment group includes subjects with a combination of ratings that fall into the lower left quadrant of Figure 1, and the remaining subjects are assigned to the comparison group). Note that this discussion generally applies to any scenario in which there is one treatment condition and one comparison condition and in which assignment is determined by two ratings, regardless of which quadrants are assigned to the treatment and comparison conditions.

In a two-rating RDD, there are two treatment-assignment frontiers — one for each rating. The rating 1 frontier, indicated by F_{R1} in Figure 1, divides the treatment and comparison groups in the upper left and upper right quadrants. It is defined for $R1 = c1$ and $R2 \geq c2$. The rating 2 frontier (F_{R2}) is defined for $R2 = c2$ and $R1 \geq c1$ and divides the treatment and comparison groups in the lower right and upper right quadrants. For this design, we define treatment-effect parameters along each of the two frontiers using the Neyman-Rubin framework as follows.

$$\psi_{R1}^{ATF} \equiv E(Y_1 - Y_0 | R1 = c1, R2 \geq c2), \text{ and} \quad (1)$$

$$\psi_{R2}^{ATF} \equiv E(Y_1 - Y_0 | R2 = c2, R1 \geq c1). \quad (2)$$

In the example with $R1$ as a reading test score and $R2$ as a math test score, $\psi_{R1}^{ATF}$is the average treatment effect at the frontier (*ATF*) for those who pass the math test but just barely pass or fail the reading test, and ψ_{R2}^{ATF} is the average treatment effect for those who pass the reading test but just barely pass or fail the math test.[5]

There are other treatment-effect parameters that may be estimated in MRRDDs. One may consider combining the frontier-specific effects (ψ_{R1}^{ATF} and ψ_{R2}^{ATF} in a 2RRDD) into a single average treatment effect at the combined frontiers, ψ^{ATF}. When the average treatment effects at the two rating frontiers are equal, then $\psi_{R1}^{ATF} = \psi_{R2}^{ATF} = \psi^{ATF}$. When $\psi_{R1}^{ATF} \neq \psi_{R2}^{ATF}$, however, ψ^{ATF} is ill defined. Although it may seem that ψ^{ATF} can be defined as a weighted average of ψ_{R1}^{ATF} and ψ_{R2}^{ATF}, Wong, Steiner, and Cook (2013) note that the weights involved should depend on the probability distribution of the ratings near the separate frontiers, and these weights are sensitive to the (arbitrary) choice of how the ratings are scaled. Standardizing ratings or using rank-ordered rating scores does not solve this problem, because the a priori validity of these scale definitions is no greater than alternative scale definitions. Consequently, our simulations focus on ψ^{ATF} only for situations in which the average treatment effect is the same at both frontiers.

The treatment-effect parameters defined by Equations 1 and 2 represent average effects of assignment to treatment based solely on the *values of one's ratings*. Hence, they represent average effects of the intent-to-treat (ITT) at a frontier. If not all sample members comply with the treatment assignment determined by their rating score, so-called "fuzzy" RDD estimation may be used to compute the average effect of receiving the treatment (rather than the average

[5] We can generalize these parameters to MRRDDs with more than two ratings by defining ratings $R1, \ldots, RJ$ and defining an average treatment effect for the k^{th} (such that $k \in (1, \ldots, J)$) frontier

$$\psi_{Rk}^{ATF} \equiv E(Y_1 - Y_0 | Rk = ck, Rj \geq cj \; for \; all \; j \neq k).$$

intent-to-treat effect). For simplicity, however, we focus on estimating "sharp" RDD parameters, in which all sample members comply with their rating-based treatment assignment (Imbens and Lemieux, 2008a).[6]

MRRDD models may also be used to investigate whether and how treatment effects and ratings are related. For example, in the 2RRDD case, one may be interested in how treatment effects at the rating 1 frontier covary with values of rating 2, or how treatment effects at the rating 2 frontier covary with values of rating 1. Such heterogeneity parameters can be examined by using interaction terms in regression models. An evaluation of how well the methods we examine estimate such treatment-effect covariation parameters is beyond the scope of the present paper. For a discussion see Reardon and Robinson (2012).

The fact that the parameters in an RDD are defined at rating frontiers is often seen as a major limitation of the generalizability of RDD findings (e.g., Imbens and Lemieux, 2008a). However, this limitation only applies to cases where treatment effects vary widely across the target population and covary strongly with observed ratings (Bloom, 2012; Lee and Lemieux, 2010). The little empirical research that exists on this issue (which is based on the evaluation of an employment program, not an education program) does not show much evidence of treatment effect-rating score covariance (Black, Galdo, and Smith, 2007; Bloom, 2012; Lee and Lemieux, 2010).

2.3 Conditions for Identification

The following conditions must be met in order for a single-rating RDD or an MRRDD to identify an average treatment effect at a frontier. For further discussion see Wong, Steiner, and Cook (2013), Reardon and Robinson (2012), and Papay, Willett, and Murnane (2011).

- *Continuity of potential outcomes at the frontier:* The functional relationships between both mean treated and untreated outcomes ($E(Y_1)$ and $E(Y_0)$) and the rating that defines a frontier are continuous at the frontier.
- *Discontinuity in the probability of treatment at the frontier:* The probability of treatment is continuous near the frontier on both sides, but discontinuous at the frontier. In the case of a "sharp" RDD, the probability of treatment is 1 on one side of the frontier and 0 on the other.

[6] The use of the terms fuzzy and sharp to describe RDD parameters should not be confused with the name for the MRRDD *estimation method* called fuzzy IV. This method can be used to estimate sharp or fuzzy MRRDD parameters. The potential for this confusion does not exist for a single RDD design.

- *Positivity, or support, at the frontier:* In order to observe a discontinuity at a frontier for an outcome of interest or for a probability of treatment, there must be a continuous distribution of potential sample members across the frontier.

- *Independent determination of individuals' ratings and the cut-point which defines the frontier:* The location of the cut-point values does not influence individuals' rating scores and individuals' rating scores do not affect decisions about where to set the cut-point values.

3. MRRDD Estimation Methods

When estimating average treatment effects at an RDD frontier, one might be inclined to select a set of sample points that are extremely close to the frontier. That is, viewing the RDD as a localized RCT, one might estimate a local average treatment effect as the simple difference between mean outcomes for sample members on the treated and untreated sides of the frontier. However, without a large number of such sample members, this estimator will be imprecise. On the other hand, including sample members who are not close to the frontier can bias this simple estimator.

Because it is typically necessary to use sample points that are not extremely close to a frontier, single-rating RDD and MRRDD treatment-effect estimators must rely on modeling the relationship between sample members' outcomes and ratings. If the functional forms used by these estimators are correctly specified, they will be unbiased. If the functional forms are not correctly specified, bias can result. The challenge here is that one does not know the correct functional forms to use.

Sharp MRRDD estimators are based on regression models of the form:

$$Y_i = f(R1_i, \dots, RJ_i, T_i) + \epsilon_i \,, \tag{3}$$

where $R1, \dots, RJ \in D \subset S$, and D is the domain of observations used to estimate the model, S is the domain of observations in the full sample, and T is a zero/one treatment-assignment indicator (Reardon and Robinson, 2012). Baseline covariates are not necessary for identification but can be included to improve the precision of treatment-effect estimates (Imbens and Lemieux, 2008a).

MRRDD estimators differ in terms of (1) the breadth of the domain of sample observations they use (*D*), (2) the nature of the function (called a "kernel") that weights these observations, and (3) the functional form (f) used to model the relationship between sample members' outcomes and ratings. As noted above, the choices of D and F involve a trade-off between bias and precision. Optimizing this trade-off correctly is difficult for a single-rating RDD and even more difficult for a MRRDD.

As noted earlier, four main methods have been used to date to estimate average treatment effects from an MRRDD (Hahn, Todd, and Van Der Klaauw, 2001; Kane, 2003; Martorell, 2005; Papay, Willett, and Murnane, 2011; Reardon and Robinson, 2012; Robinson, 2008, 2011; Wong, Steiner, and Cook, 2013). These methods are described in detail below.

1. ***The surface method:*** The surface method, or the multivariate method, models the multidimensional response surface that represents the relationship between sample members' outcomes and ratings (Papay, Willett, and Murnane, 2011; Reardon and Robinson, 2012; Wong, Steiner, and Cook, 2013). Hence, the functional form, f, of this surface includes all ratings that are involved in treatment assignment, and D covers all quadrants defined by the ratings.[7] Because this method models the response surface across all quadrants, it simultaneously estimates the parameters defined at each treatment frontier. In the case of a 2RRDD this implies simultaneously estimating ψ_{R1}^{ATF} and ψ_{R2}^{ATF}.

2. ***The frontier method:*** This method reduces the MRRDD to a single-rating RDD by limiting the analysis to sample members whose treatment assignment is determined by only one rating. It then uses the relationship between these sample members' outcomes and the single rating as the basis of a standard, single-rating RDD (Jacob and Lefgren, 2004; Kane, 2003; Reardon and Robinson, 2012; Wong, Steiner, and Cook, 2013). For example, to estimate ψ_{R1}^{ATF}, one may restrict D to those with $R2 \geq c2$ and fit regression models that specify outcomes as a function of $R1$ plus a treatment-assignment indicator, T.[8] This process can be repeated to estimate ψ_{R2}^{ATF} by restricting D to those with $R1 \geq c1$ and fitting a regression model of outcomes on $R2$ and the treatment indicator. Partitioning the sample in this way may reduce precision relative to the surface method because it restricts D.[9]

3. ***The fuzzy IV method:*** This method uses instrumental variables (IV) analysis in a way that allows one to use all available sample points to estimate a treatment effect at a single frontier by accounting for "noncompliance" caused by treatment assign-

[7] Technically, although the domain for this method must include some samples from all quadrants, it does not necessarily require using all sample members.

[8] Note that $R2$ and other baseline covariates could be added to this regression to improve precision.

[9] As described later, the number of sample members omitted for this method and its corresponding loss of precision depends on the location of cut-points for the ratings.

ment based on the other rating (or other ratings) (Reardon and Robinson, 2012; Wong, Steiner, and Cook, 2013). For example, ψ_{R1}^{ATF} (which, recall, is an ITT parameter, or a sharp RDD parameter) in a 2RRDD can be estimated by defining an indicator variable $z1$ as an instrument for treatment assignment, *T*, such that $z1 = I(R1 < c1)$, and implementing IV analysis to estimate the mean effect of treatment assignment *based on the value of* $R1$. This approach accounts for the fact that treatment assignment based on rating 1 is "overridden" for some sample members by treatment assignment based on other ratings. Thus, even though treatment assignment is fully determined by sample members' ratings, compliance with treatment assignment based on a single rating is incomplete.[10] Compared with the frontier method, this method uses more sample points. However, the added uncertainty of instrumental variables analysis can offset the precision gain produced by its increased sample size (Reardon and Robinson, 2012).

4. ***The binding-score method:*** This method, sometimes referred to as the centering method, entails collapsing multiple ratings into a single rating and estimating an average treatment effect for the entire study sample using a single-rating RDD (Martorell, 2005; Reardon and Robinson, 2012; Robinson, 2011). In Figure 1, where only the sample members who fall below the cut-point value for one or both ratings are assigned to the treatment, the binding score is defined as the *minimum value* of the two ratings. To use this approach it is necessary to center the rating scores at their treatment assignment cut-point values. It may also be useful, though it is not necessary, to express the two ratings in comparable units by standardizing each as a z-score centered on its cut-point. The sample domain for the binding-score method is not restricted and its functional form can be a complex function of the original ratings. The major drawback of this approach is that it only provides an estimate of the overall average treatment effect (ψ^{ATF}); it does not provide estimates of the frontier-specific effects (ψ_{R1}^{ATF} or ψ_{R2}^{ATF}). Thus only when treatment effects are homogeneous across frontiers does the binding-score method yield findings that are comparable to those provided by the other methods. Moreover, because ψ^{ATF} has a clear interpretation only if $\psi_{R1}^{ATF} = \psi_{R2}^{ATF}$, the binding-score method should be used only when this assumption holds as a reasonable approximation.

The present paper compares the preceding four estimation methods using simulations that allow us to compute the bias and variance of each estimator when applied to the same data.

[10] In Figure 1, individuals in the lower right quadrant would be noncompliers because $z1$ implies that they should not have received the treatment. This form of incomplete compliance is the "fuzziness" that is addressed by the IV analysis for an MRRDD.

This work builds on a simulation study by Wong, Steiner, and Cook (2013), which also computes the bias and variance of the four estimation methods when they are applied to the same two-rating design. Their simulations focus on the performance of the estimation methods when two factors are varied: (1) the complexity of the true data-generating response surface with respect to homogeneous versus heterogeneous treatment effects, and (2) the metric and scale of the two ratings. Wong, Steiner, and Cook (2013) vary the metric and scale of the ratings in order to assess whether standardizing rating variables allows one to estimate ψ^{ATF}, and demonstrate that it does not, because, as we discuss above, ψ^{ATF} has a clear interpretation only if $\psi^{ATF}_{R1} = \psi^{ATF}_{R2}$.

We build on the work of Wong, Steiner, and Cook (2013) by varying additional factors that a researcher would encounter in practice — the correlation between ratings, the locations of the cut-points for the ratings, and the functional forms of the relationships between the outcome and the ratings (i.e., quadratic versus linear). Moreover, when comparing the four methods, we disentangle the implications of how the methods perform *in theory* from how they are likely to perform when implemented *in practice*. In particular, we advance previous work by examining how the optimal bandwidth selection approach used to implement the estimation methods in practice influences statistical inferences, an important issue which has not yet been explored.

4. Our Simulations

Our simulations were based on values for $R1$ and $R2$ that were generated from a bivariate normal distribution with a mean of 0 and a standard deviation of 1 for each rating. Simulations were conducted for three correlations between the ratings (0.20, 0.50, and 0.90) and for three combinations of cut-points (the 50th percentile values for both ratings, the 30th percentile values for both ratings, and the 30th percentile value for one rating and the 70th percentile value for the other rating). All ratings were centered around their cut-point values.

Outcomes were generated based on ratings and treatment assignment using the following four models, which represent a pattern of increasing complexity:[11]

Model 1: $$Y_i = 0.4T_i + 0.5R1_i + R2_i + \epsilon_i \tag{4}$$

Model 2: $$Y_i = 0.4T_i + 0.5R1_i + R2_i + 2R1_i^2 + R2_i^2 + \epsilon_i \tag{5}$$

Model 3: $$\begin{aligned} Y_i = 0.4T_i + 0.5R1_i + R2_i + 2R1_i^2 + R2_i^2 - 0.1T_iR1_i \\ - 0.2T_iR2_i + \epsilon_i \end{aligned} \tag{6}$$

[11] Model 1 is similar to the simplest model used by Wong, Steiner, and Cook (2013).

Model 4: $$Y_i = 0.4T_i + 0.5R1_i + R2_i + 2R1_i^2 + R2_i^2 - 0.1T_iR1_i - 0.2T_iR2_i - 0.08T_iR1_i^2 - 0.08T_iR2_i^2 + \epsilon_i \quad (7)$$

where $\epsilon \sim N(0,1)$.

Models 1 and 2 specify a constant treatment effect of 0.4 (the coefficient on the treatment indicator, T_i) at both frontiers ($\psi_{R1}^{ATF} = \psi_{R2}^{ATF} = \psi^{ATF} = 0.4$). Note that Model 1 is linear in the ratings whereas Model 2 is quadratic in the ratings. Also note that the relative magnitudes of the coefficients on the quadratic terms in Model 2 produce substantial curvature, which in turn provides a substantial challenge for the estimation methods being examined.

Models 3 and 4 specify heterogeneous treatment effects that depend on the values of the ratings (Model 3 is linear in the ratings, and Model 4 is highly nonlinear). The values of ψ_{R1}^{ATF} and ψ_{R2}^{ATF} are not the same for the data generated by these models. In addition, these values cannot readily be computed from the model coefficients because the joint density distribution of rating scores is not uniform along the frontiers. Consequently, the average treatment effect must be computed by integrating the "marginal" treatment effect over the density along each frontier (see Appendix A).

The findings reported in this paper are based on 20 simulated scenarios. Holding the cut-point percentiles at 50/50, we varied the correlations between the ratings (0.20, 0.50, and 0.90), and holding the correlation between the ratings at 0.20, we varied the cut-point percentiles (50/50, 30/30, and 30/70). We did this for each of the four outcome response models (Equations 4-7).

5. Implementation of the Four MRRDD Estimation Methods

The first goal of our simulation analysis was to evaluate the MRRDD estimation methods in theory, or under "ideal conditions" with full information about the data-generating mechanisms. Appendix B describes the implementation of the four methods using the known functional form of the data generation models.

In practice under "normal operating conditions," the functional forms of the data-generating models are unknown. For example, the functional form of the surface response model, $f(R1, R2, T)$, is unknown to the analyst. Therefore, analysts are faced with the dual challenge of choosing an estimation method and choosing the best way to implement it based on information available from a study sample. Recall from above that all of our 2RRDD estimators have the form

$$Y_i = f(R1_i, R2_i, T_i) + \epsilon_i, \quad (8)$$

where $R1, R2 \in D \subset S$, such that D is the domain of observations used to estimate a model and S is the domain of all observations in a study sample. As discussed above, the goal when selecting a functional form, f, and domain, D, for a given estimation method is to optimize the bias-precision trade-off. An analyst might approach making decisions about f and D in several ways. For example, she might compare plots of data corresponding to different functional forms for different sample domains to determine the best model fits.

Combining such a visual assessment with the computational assessment of goodness-of-fit may guide decisions about an optimal combination of f and D. However, with this approach, it is not possible to take uncertainty of bandwidth selection into account when making statistical inferences. Alternatively, an analyst could take a more systematic and prespecified approach and rely on a programmed algorithm that selects the domain, D, for a specified functional form, f, such that f approximates the data within the chosen domain as well as possible.[12] We developed such an algorithm for three of the four MRRDD estimation methods: the frontier method, the fuzzy IV method, and the binding-score method.[13] For each method, we specified a linear functional form and programmed the algorithm to choose a domain that minimized the mean squared error of a corresponding estimator implied by the method. The three algorithms are variants of a conventional bandwidth selection approach using leave-one-out cross-validation that is currently employed for local linear regression models in single-rating RDDs (e.g., Lee and Lemieux, 2010).[14] A step-by-step description of each algorithm is presented in Appendix D.

We estimated effects using two different linear regression models fit to the data within the optimal bandwidth determined by each algorithm corresponding to each estimation method. The first model specifies outcomes as a linear function of $R1$, T, and an interaction between $R1$

[12] Alternatively, one could jointly select an optimal D *and* f. For example, one candidate for f might be a simple linear regression; another might also include a quadratic term for the rating corresponding to the frontier of interest; and a third might include quadratic terms for both ratings. The influence on precision of using more complex models is mixed, as a more complex model involving more higher-order polynomial terms may lead to precision loss unless it allows using much more data.

[13] We did not implement a bandwidth selection algorithm for the surface method because of the major challenges of doing so. First, for a 2RRDD, the surface method requires choosing an optimal two-dimensional domain instead of a simpler one-dimensional bandwidth; and, among other complications, it is not clear what the shape of this domain should be. Second, the leave-one-out cross-validation algorithm described by Ludwig and Miller (2007) and Imbens and Lemieux (2008a) that we use for the other MRRDD estimation methods requires identifying the sample points that are nearest to the frontier of interest. However, when there are multiple frontiers, choices about the scale of each rating will affect decisions about which points are nearest to the frontiers, making bandwidth selection sensitive to scaling choices.

[14] We could have also used Imbens and Kalyanaraman's (2009) approach for optimal bandwidth selection (as Wong, Steiner, and Cook, 2013, do), but we chose leave-one-out cross-validation for ease of presentation and programming and because we did not think it would change the conclusions of the analysis.

and T, thereby allowing different slopes for $R1$ on the right and left sides of the frontier. The other model adds $R2$ as a baseline covariate in order to improve precision with a common coefficient on both sides of the frontier.[15] Having chosen an optimal bandwidth for a linear functional form for each of the three methods, we then used each method to estimate the average treatment effect at the frontier for rating 1 (ψ_{R1}^{ATF}) and at the frontier for rating 2 (ψ_{R2}^{ATF}). This section describes the specific regression models used for each method and how ψ_{R1}^{ATF} was estimated for each method. To estimate ψ_{R2}^{ATF} we reversed the roles of $R1$ and $R2$. This process is described in more detail below.

The frontier method. To use the frontier method to estimate ψ_{R1}^{ATF} we first chose all sample members for whom $R2 \geq 0$. We then selected all of the remaining sample members who were inside the optimal bandwidth on the left or right side of the rating 1 frontier. For the resulting subsample, we fit one of the following two models:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 R1_i + \beta_3 R1_i T_i + \epsilon_i, \text{ or} \qquad (9)$$

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 R1_i + \beta_3 R1_i T_i + \beta_4 R2_i + \epsilon_i. \qquad (10)$$

The model in Equation 10 includes rating 2 as a covariate to improve precision of the estimator of ψ_{R1}^{ATF}, which is equal to $\hat{\beta}_1$ in both models.

The fuzzy IV method. We identified all sample members in the smaller of the optimal bandwidths for the first-stage and reduced-form models and estimated the following models:

First-stage: $$T_i = \gamma_0 + \gamma_1(1 - z1_i) + \gamma_2 R1_i + \gamma_3(1 - z1_i)R1_i + e_i, \qquad (11)$$

Reduced-form: $$Y_i = \beta_0 + \beta_1(1 - z1_i) + \beta_2 R1_i + \beta_3(1 - z1_i)R1_i + \epsilon_i, \qquad (12)$$

where $z1 = 0$ on the right side of the frontier and $z1 = 1$ on the left side of the frontier.[16] To improve precision, we also fit the following models, which add $R2$ as a covariate:

First-stage: $$Y_i = \gamma_0 + \gamma_1(1 - z1_i) + \gamma_2 R1_i + \gamma_3(1 - z1_i)R1_i + \gamma_4 R2_i + e_i, \qquad (13)$$

Reduced-form: $$Y_i = \beta_0 + \beta_1(1 - z1_i) + \beta_2 R1_i + \beta_3(1 - z1_i)R1_i + \beta_4 R2_i + \epsilon_i. \qquad (14)$$

[15] Note that this model is different from the model used to choose an optimal bandwidth, which does not include $R2$ as a covariate. Also note that we constrain the coefficient for $R2$ to be the same on both sides of the frontier because it is added only to improve precision; allowing its coefficient to vary would require using numeric integration to estimate a mean treatment effect at the $R1$ frontier, which would greatly increase the complexity of the analysis and is an approach that is not widely used by applied researchers.

[16] Therefore, in Equations 13 and 14, $(1 - z1)$ is the “instrument.”

For both models (without or with the $R2$ covariate), we then computed the following Wald estimator of the mean treatment effect at the frontier for rating 1:

$$\hat{\psi}_{R1}^{ATF}(fuzzyIV) = \frac{\hat{\beta}_1}{\hat{\gamma}_1}. \tag{15}$$

The binding-score method. We identified all sample members who were in the optimal bandwidth and then fit the following model to data for these sample members:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 R_{BS_i} + \beta_3 T_i R_{BS_i} + \epsilon_i, \tag{16}$$

where $R_{BS} \equiv \min(R1, R2)$. The resulting value for $\hat{\beta}_1$ is our estimate of $\psi_{R1}^{ATF} = \psi_{R2}^{ATF} = \psi^{ATF}$. To increase precision we added the original ratings ($R1$ and $R2$) as covariates to Equation 16 and estimated the following model:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 R_{BS_i} + \beta_3 T_i R_{BS_i} + \beta_4 R1_i + \beta_5 R2_i + \epsilon_i. \tag{17}$$

The resulting value for $\hat{\beta}_1$ is our estimate of $\psi_{R1}^{ATF} = \psi_{R2}^{ATF} = \psi^{ATF}$.

6. Simulation Results

Our simulation results are for four data-generating models (Equations 4-7), plus the three correlations between $R1$ and $R2$ (0.2, 0.5, and 0.9) and the three combinations of cut-point percentiles for $R1$ and $R2$ (50/50, 30/70, and 30/30) noted earlier. For specified combinations of these conditions, we simulated 500 samples of size 5,000 for analyses under normal conditions in practice without knowledge of the data-generating models, and 500 samples of size 1,000 for analyses under the theoretical condition of full information about the data-generating models. For each simulation, we computed the bias, variance, and mean squared error of the estimators. We also computed the relative variance and relative mean squared error of one estimator with respect to another estimator by taking their ratios.

Appendix B presents results for analyses that use knowledge of the underlying data-generating models. As expected, all three estimation methods produced unbiased estimates under all simulated scenarios. In terms of precision, the surface method is always superior to the others. For example, the surface method is more precise than the frontier method because the frontier method uses data for only part of a given sample whereas the surface method can use data for the whole sample. In addition, the surface method is more precise than the fuzzy IV method because of the added uncertainty produced by the two-stage IV estimation process. Indeed, the fuzzy IV method was always the least precise of the methods examined. Thus even though fuzzy IV uses data for all sample points and therefore could theoretically be more precise than the frontier method, which uses data for only a portion of the sample points

(Reardon and Robinson, 2012), this was never the case in our simulations. Note that Wong, Steiner, and Cook (2013) find the same result.

Turning to results from estimation using local linear regression techniques, bandwidth selection results are presented in Appendix D. Overall, we saw that a wide range of bandwidths were selected for all of the data-generating models and for each of estimation methods. However, larger bandwidths tend to be selected more frequently when the data are generated by a linear model. This reflects the fact that a linear estimation model is the correct fit for data produced by a linear data-generating model, even for sample points that are far away from the cut-points for each rating. In contrast, the fuzzy IV method tends to use bandwidths that are smaller than those used by the other two methods. Finally, very small bandwidths are rarely selected. This suggests that moderate to large bandwidths can reasonably approximate even highly nonlinear data-generating models.

Figures 2-5 present, for each data-generating model, visual comparisons of the bias and variance of the effect estimators — across simulated scenarios, across estimation methods, and across models that do not or do include the other rating as a covariate.[17] Recall that the models that do include the other rating are not fully consistent with the models used to select optimal bandwidths. Consequently, the selected bandwidths may not be optimal for the functional form that is used for estimation, and some bias could result. The circles in these plots indicate average effect estimates across 500 simulated samples. The distances of these circles from the horizontal lines illustrate the biases of the corresponding effect estimators. The vertical lines in these plots span from the 2.75th percentile to the 97.5th percentile of the 500 effect estimates. Therefore, the lengths of these vertical lines — the empirical confidence intervals — show the variability of the effect estimators. Tables 1-4 present detailed summary statistics that correspond to the results shown in Figures 2-5 when the other rating is included as a covariate in the final estimation model. (Tables with results without the other rating included as a covariate are available in Appendix E.)

When the data-generating model is linear, all estimators are unbiased. When the data-generating model is nonlinear, we do observe bias in some cases. For example, for the frontier method we observe a statistically significant bias in 10 percent of the cases for Models 2 and 4 when we do not include a covariate and in 65 percent of the cases when we do include a covariate. However, these biases are typically small.

[17] Recall that when estimating effects at the rating 1 frontier, $R2$ is used as a covariate for the frontier and fuzzy IV methods, whereas $R1$ and $R1$ are used as covariates for the binding-score method.

For Model 2, the binding-score method yields statistically significant bias in 20 percent of the simulated scenarios when a covariate is not included and in 40 percent of the cases when a covariate is included, but this bias is always small. In contrast, the fuzzy IV method yields statistically significant and substantial bias (sometimes more than 100 percent) in 45 percent of the cases under Models 2 and 4 when a covariate is not included and in 65 percent of the cases when a covariate is included. Thus the fuzzy IV method is prone to bias more than the other two methods. Although we tend to see bias more often when we include a covariate, the gains in precision produced by including the covariate usually offset the bias that it creates. Thus adding a covariate typically reduces the mean squared error.

Tables 1-4 suggest that the relative performance of the binding-score estimator is more robust than the frontier method; it has the lowest mean squared error for most simulated scenarios under Models 1 and 2 (homogeneous treatment effects).[18] The binding-score estimator outperforms the frontier estimator more frequently and by a larger margin when the correlation between the ratings is high. Conceptually, this is because a high correlation between the ratings implies that there is often relatively little data on one side of a frontier for estimation by the frontier method. As the correlation between ratings approaches 1, however, the binding-score method becomes identical to a single-rating RDD, which uses all of the data for a given sample. Finally, note that the fuzzy IV estimator is the least robust. Not only is it the most likely to exhibit bias (and much more substantial bias), but it also has higher variances than the other estimators, often by a substantial margin. Thus its mean squared errors are the largest by far.

Tables 1-4 also provide information about the implications for statistical inference of our combined local linear regression/optimal bandwidth approaches.[19] We address this question by presenting two separate estimates of the variance of treatment effect estimates. The first estimate, reported in the column labeled $Var(\hat{\psi})$, is an approximation of the true variance of the effect estimator. We obtained this estimate by computing the variance of $\hat{\psi}_{R1}^{ATF}$across the 500 simulated samples. This variance estimate takes into account the uncertainty introduced by optimal bandwidth selection as well as that due to actual estimation of the local linear regression. The second estimate of the variance for impact estimate, reported in the column labeled "Mean of Estimate of $(\hat{\psi})$," is the mean OLS or IV estimated variance from each run of 500 simulated samples. This represents the estimated variance one would get if the uncertainty due to the bandwidth selection process were ignored.

[18] Recall that we did not implement the binding-score estimator for simulations of Models 3 and 4 because this estimator does not identify a meaningful parameter when treatment effects vary with ratings.

[19] These issues apply with equal force to single-rating RDDs.

Tables 1-4 suggest that the true variances of our impact estimators are typically larger than their estimated values. These differences are small, however, and therefore, at least for the range of conditions represented by our simulations, the variance estimates that do not account for the uncertainty added by optimal bandwidth selection are typically good estimates of the true variance of the impact estimators. We explore this issue further in our application study.

7. Application Study

This section walks through the steps needed to apply the lessons learned from the preceding simulations to an analysis of existing data for a 2RRDD. Specifically, the section describes the data that were used, defines the parameters, outlines the steps that were taken to estimate these parameters, and interprets the findings.

7.1 Data

Data for the present analysis were used by Robinson (2011) to study the effects of changes in status, instructional services, and educational settings that occur when students in a large urban district were reclassified from English learner (EL) to reclassified fluent-English-proficient (R-FEP). To be reclassified, ELs in this district must score above specified thresholds on five standardized tests or subtests: the California English Language Development Test (CELDT) overall, the CELDT reading, writing, and listening/speaking subtests, and the California Standards Test of English Language Arts (CST ELA). ELs who exceed all five thresholds become eligible for reclassification; the final decision about this determination is based on teacher judgments and parent consultations. Robinson (2011) used binding-score RDDs with IV estimation to study the effects of reclassification on the marginal ELs' future academic achievement, course taking, and school attendance.

We use the same data in order to illustrate how to apply the methods described to a real-world data set. To do so, we first had to reduce the five assessment scores (five RDD ratings) that are actually used to reclassify ELs to two assessment scores that can serve as the ratings for a 2RRDD. This was accomplished by limiting our analysis sample to ELs who scored above the threshold for three of the five assessments: the CELDT overall, the CELDT writing subtest, and the CELDT listening/speaking subtest. This produced the largest possible analysis sample for which eligibility for reclassification was determined solely by two assessments: the CELDT reading subtest and the CST ELA.[20] This subsample constitutes the equivalent of a 2RRDD in which ELs who pass both of the remaining assessments (the two RDD ratings) are eligible for

[20] We also selected the set of grade levels that would maximize our sample size. The sample used by Robinson (2011) includes elementary and high school students as well.

reclassification (the treatment condition) and those who do not pass both of the assessments are not eligible for reclassification (the control condition). The resulting 2RRDD has a sample of 6,102 middle-school students (sixth-, seventh-, and eighth-graders) and two ratings that determine their treatment or control status.[21]

The estimated correlation between the two ratings (students' scores on the CST ELA and the CELDT reading subtest) is 0.39 in our sample, which is well within the range of scenarios represented by our simulations. About 53 percent of our sample members scored above the threshold for the CST ELA rating, and about 75 percent scored above the threshold for the CELDT reading subtest rating, which also is well within the range of scenarios represented by our simulations. As in our simulations, we centered the values of each rating around its cut-point, so that both cut-points equal zero, and we standardized the values of each rating by dividing them by the standard deviation of the rating for each sample member's grade.

7.2 Parameters

Our analyses differ in the following ways from those reported by Robinson (2011): (1) We restrict our analyses to students passing three of the five thresholds (as noted above); (2) we do not estimate the actual effects of reclassification, but rather estimate ITT effects of passing the two remaining thresholds (i.e., the effect of becoming eligible for reclassification); (3) we examine only one outcome (the CST ELA post-test); and (4) we estimate effects on students in grades six through eight only. That is, for the purpose of illustrating estimation methods for a 2RRDD, we estimate the following:

1. The effect of just barely passing or failing the CST ELA threshold among those who passed the CELDT reading threshold. This parameter is defined as

$$\psi_{CST\ ELA}^{ATF} \equiv E(Y_1 - Y_0 | CST\ ELA_s = 0, CELDT\ reading_s > 0), \qquad (18)$$

   where Y_1 for a given student is the value of the potential outcome if she exceeds the CST ELA threshold and Y_0 is the potential outcome if she does not exceed this threshold. $CST\ ELA_s$ and $CELDT\ reading_s$ are the ratings which were standardized as described above.

2. The effect of just barely passing or failing the CELDT reading subtest threshold among those who passed the CST ELA threshold. This parameter is given by

[21] As in Robinson (2011), all students in our sample have three years of consecutive assessment data.

$$\psi_{CELDT\ reading}^{ATF} \equiv E(Y_1 - Y_0 | CELDT\ reading_s = 0, CST\ ELA_s > 0). \quad (19)$$

7.3 Estimation

We approach estimation the same way we did when mimicking estimation in practice in our simulation study — using local linear regression. Thus for each effect parameter and estimation method, we first apply our algorithm to select an optimal bandwidth for a linear functional form. We then estimate the effect parameter at its corresponding frontier, using data within the selected bandwidth.[22]

The first step in this process is to decide which of the 2RRDD estimation methods to use. Although our simulation study indicates that the surface estimation method has the potential to be the most precise, regardless of the correlation between ratings or the location of their cut-points, implementing the surface method with an optimal bandwidth strategy is not feasible at this time for the reasons that were discussed in a footnote earlier.[23] Therefore, we do not consider surface estimation for this application study.

Recall that the fuzzy RDD method performed least well in all of our simulated scenarios. Thus even though under full information it is theoretically possible for the fuzzy IV method to be more precise than the frontier method (because the former method uses more data than does the latter), we never observed this result in our simulations. Moreover, in practice, when the functional forms of the data-generating models are not known, the fuzzy IV estimator was frequently and substantially biased. Thus, we do not recommend using this method and do not use it for the present application.

This leaves the frontier and binding-score methods to consider. Recall that our simulations indicate that the frontier method is probably the best approach for estimating separate treatment effects at each of two frontiers. It is relatively straightforward to implement, and the approach works well when using local linear regression. On the other hand, the frontier method requires a large number of sample points around both frontiers of interest. In addition, estimat-

[22] There are other approaches that one could use to deal with functional forms for 2RRDDs, like examining the apparent functional form visually through plots of the data. However, the focus of this paper is on how to select among the four basic 2RRDD estimation methods and how to implement the selected method when using local linear regression.

[23] This point of view differs from that of Papay, Willett, and Murnane (2011), who jointly select bandwidths in two dimensions that constitute a single, rectangular area bounded by selected values for rating 1 (in one dimension) and rating 2 (in the other dimension). This approach may alter the parameter, however, because it excludes points near one threshold that are far from the other threshold.

ing treatment effects at each frontier separately involves testing multiple hypotheses, which can require statistical adjustments that further erode precision.

Our simulations also indicate that whenever it is reasonable to assume that the average program effects at the two frontiers are approximately equal to each other, the binding-score method should be preferred over the frontier method. This is because given homogeneity of effects across frontiers, the binding-score method had the lowest mean squared errors. However, when treatment effects differ appreciably across frontiers, the binding-score method can only produce a weighted average of these effects, with weights that are arbitrary (and have no inherent meaning) because they depend on the scale or distribution of the ratings involved. Since it is difficult to test the homogeneity of intervention effects across frontiers without revealing the actual values of the resulting estimates (and thereby potentially biasing researchers' selection of an estimation method), we recommend proceeding as follows. If a researcher has a plausible theory about why treatment effects should be approximately homogeneous across frontiers, she should: (1) present the theoretical rationale for expecting homogeneity; (2) use the binding-score method to maximize precision; (3) use the frontier method to estimate a separate treatment effect at each frontier; and (4) report both sets of estimates.

7.4 Implementation of the Methods

To compute effect estimates, we implemented both the frontier and binding-score methods the same way we did in our simulation study. The only exception is that instead of defining bandwidths as having equal width on each side of the cut-point, our algorithm allowed for different widths on each side. We made this choice to be more flexible because plots of the data suggested that the relationship between the outcome and the ratings varied on different sides of the cut-points.

To compute statistical inference for the effect estimates, we implemented a nonparametric bootstrap. Recall that in our simulation study, bandwidth selection added some uncertainty to our 2RRDD effect estimators. The added uncertainty led to small differences between the true standard errors and standard errors estimated by OLS. However, it is possible that data-generating distributions that differ from those we simulated could result in larger discrepancies. The nonparametric bootstrap mimics our simulations by repeatedly sampling (with replacement) from our applied data set.[24] Within each bootstrap sample, for each effect parameter, we repeated the entire estimation process (just as we did in each simulation sample), including

[24] When using the frontier estimation method, we resampled after subsetting the data. That is, we resampled observations for which $CST\ ELA_s > 0$ when estimating $\psi^{ATF}_{CELDT\ reading}$, and we resampled observations for which $CELDT\ reading_s > 0$ when estimating $\psi^{ATF}_{CST\ ELA}$.

optimal bandwidth selection. We used 3,000 bootstrap samples. The standard deviation of the resulting 3,000 effect estimates approximates the standard error of each estimate, which includes the uncertainty of bandwidth selection. We then compared this with the OLS approximation of the standard error, which does not include the uncertainty of bandwidth selection.

7.5 Results

The results from both methods are presented in Table 5. The first three columns of the table present information about the optimal bandwidth that was selected: $\%bw_l$ is the percent of the data to the left of the frontier (with a rating value less than zero) that is included in the bandwidth; $\%bw_r$ is the percent of the data to the right of the frontier (with a rating value at or greater than zero) that is included in the bandwidth; and N_{bw} is the total number of observations in the bandwidth and thus is used for fitting the regression model for the analysis. The rest of the columns in Table 5 present the effect estimates, the OLS estimates of the standard errors, the bootstrap estimates of the standard errors, the 95 percent confidence intervals computed from the OLS estimates of the standard errors, and the empirical 95 percent confidence intervals computed from the bootstrap estimates (i.e., the 2.5th and 97.5th percentiles in the distribution of the estimates).

When using the frontier method to estimate $\psi^{ATF}_{CST\ ELA}$, the optimal bandwidth includes 40 percent of the data to the left of the frontier and 50 percent of the data to the right of the frontier. This yields a sample of 2,115 students. When using the frontier method to estimate $\psi^{ATF}_{CELDT\ reading}$, the optimal bandwidth includes 60 percent of the data to the left of the frontier and 90 percent of the data to the right of the frontier. This results in a sample size of 2,788 students. Finally, when using the binding-score method to estimate ψ^{ATF}_{BS}, which is a weighted average of $\psi^{ATF}_{CST\ ELA}$ and $\psi^{ATF}_{CELDT\ reading}$, the optimal bandwidth includes 40 percent of the data to the left of the frontier and 60 percent of the data to the right of the frontier, resulting in a sample size of 3,004. The binding-score method has a larger sample size because it does not throw out any observations before bandwidth selection.

We recommend that analysts produce plots similar to those in Figure 6 to check that the selection of the optimal bandwidths is reasonable. Each plot is similar to a typical plot for a single-rating RDD, as both the frontier and binding-score methods reduce analyses to a single-rating RDD. All plots represent the CST ELA post-test on the *Y*-axis and the rating of interest for each analysis is on the *X*-axis. Rather than plotting all of the points, we have plotted the mean outcomes in bins of size 0.1 standard deviation of the rating. The mean outcomes are represented by the centers of the circles which are scaled to show the number of observations within the bin. Reducing the data in this way helps the visualization of the functional form of the outcome-rating relationship. In each plot, the vertical line drawn at $X = 0$ represents the frontier, and the lines on either side illustrate boundaries of the optimal bandwidth. In each of these plots, we can see that the relationship between the outcome and the rating is relatively linear within the bandwidth, while the relationship becomes less clear as we move farther away from the frontier.

Table 5 shows that the frontier method yields $\hat{\psi}^{ATF}_{CST\,ELA} = -2.2$, with an OLS standard error of 2.2 and a bootstrap standard error of 4.4. We also find $\hat{\psi}^{ATF}_{CELDT\,reading} = -1.9$, with an OLS-estimated standard error of 3.1 and a bootstrap-estimated standard error of 3.6. Also, when using the binding-score method, we find $\hat{\psi}^{ATF}_{BS} = 0.4$, with an OLS-estimated standard error of 2.1 and a bootstrap-based standard error of 2.7.

For all effect estimates, the difference between the bootstrap standard error and the OLS standard error is nonnegligible. This demonstrates that the bandwidth selection is adding uncertainty to the effect estimates, which should be taken into account. Therefore, we recommend that analysts do not rely on standard error estimates returned by statistical software. Instead they should implement a nonparametric bootstrap procedure, as described above, to account for all sources of uncertainty in the estimated standard errors.

Overall, the present analysis suggests that effect estimates from both methods should be reported and discussed. Based on these results, we cannot reject the null hypothesis that average effects are zero at both frontiers. Had the frontier method produced convincing evidence of differential effects across the two frontiers, we probably should have disregarded or substantially downplayed results from the binding-score method. On the other hand, if we had a strong theory for why average effects should be the same at both frontiers *and* if the frontier method produced convincing evidence of similar effects, then we should also report and discuss the binding-score effect estimate.

8. Discussion

The present paper compares the statistical properties of alternative 2RRDD estimation methods when they target the same causal parameters — average treatment effects at frontiers defined by a cut-point value of a single rating. Using a variety of simulated scenarios, the paper illustrates how the different methods perform both under theoretically ideal conditions when the true functional form of the 2RRDD model is known and, more important, under the conditions that occur in practice when this functional form is not known. To implement the estimation methods under these latter conditions, we use local linear regression (i.e., optimal bandwidth selection for a linear functional form), both for our simulation studies of the properties of these estimators and for our application of the methods to an existing data set for a major educational intervention.[25]

We find that each of the four methods has its own merits and limitations, depending on the structure and characteristics of existing data, especially the correlation between the two ratings and the locations of their cut-points. For example, as the correlation between the two ratings increases, the precision advantage of the binding-score method over the frontier method increases. Similarly, because the location of cut-points influences the proportion of sample points omitted by the frontier method, this factor can contribute to differences in its relative precision. Nonetheless, these factors do not affect our overall recommendations about when to use each method.

Two of the four methods have problems that severely limit their applicability. For example, the fuzzy IV method was found to be extremely imprecise for all of the data-generating scenarios that were considered. Thus it seems to have little utility for real-world application. In addition, even though the surface method is potentially the most precise, it is especially sensitive to functional form misspecification. And it is not clear how to avoid the problem by implementing the surface method using local linear regression.

The frontier and binding-score methods seem to have the best statistical properties. In addition, because both methods reduce the multiple-rating design to one or more single-rating RDDs, they are relatively simple to implement using conventional methods. Hence they appear to be the most viable candidates for real-world application.

As we did in our application study, we recommend that an analyst use the frontier method to analyze treatment effects for a 2RRDD unless there is a strong theoretical reason for assuming that treatment effects are approximately homogeneous across frontiers. When the theoretical case for homogeneity is strong, we recommend using both the frontier method and the binding-score method and presenting and comparing their results.

Finally, the uncertainty of bandwidth selection contributes to the variance of MRRDD estimators (as well as to the variance of single-rating RDD estimators). Therefore not taking this uncertainty into account, by relying on OLS or IV estimates of standard errors, can cause one to understate the total uncertainty that is involved and thereby inflate the likelihood of Type I errors. To ensure that bandwidth selection uncertainty is properly accounted for, we recommend

[25] We do not explore alternative approaches to implementing the four methods in practice because (1) the approach we use is effective and easy to apply and (2) our goal is to assess the relative strengths and weaknesses of the four estimation methods when they are implemented similarly.

a nonparametric bootstrap procedure, in which bandwidth selection is repeated for each bootstrapped sample.

Several important issues that are specific to MRRDDs need further investigation. First, while our simulations attempted to mimic a range of scenarios that could occur in practice, there may be other scenarios — other combinations of data-generating distributions, correlations, and cut-point locations — that could call for modifications to how one approaches estimation. Second, more investigation is needed to understand sample size requirements for 2RRDDs and higher dimension MRRDDs — not just for sufficient power but for the optimal bandwidth selection implementation strategy to be operational and valid. Moreover, while our simulations ensured that there were substantial data points on either side of all treatment frontiers, in many real-world MRRDDs the distribution of the points may be far less optimal — such that the data are sparse close to the frontier even if dense farther from the frontier. Therefore, the distribution of data points near the frontiers needs to be taken into account when considering sample size and power. Third, because our goal was to compare the four MRRDD estimation methods when implemented with a common, prespecified nonparametric approach (using optimal bandwidth selection), we did not explore the robustness of effect estimates to alternative specifications of bandwidth selection algorithms. While alternative specifications should not affect which of the four MRRDD estimation methods one would choose, they could influence the bias or precision of all of the estimation methods. Finally, other parameters may be estimated from MRRDDs. For example, researchers may be interested in estimating the effects of treatment receipt instead of intent-to-treat, which may require an instrumental variables methodology *in addition* to the methods described in the present paper. Researchers may also be interested in estimating treatment variation parameters. Further guidance is needed to estimate these parameters.

In summary, this paper makes valuable contributions to the literature on MRRDDs in that it makes concrete recommendations for choosing among MRRDD estimation methods, for implementing the chosen method using local linear regression, and for providing accurate statistical inferences. Our simulations and applied example focused entirely on 2RRDDs. As the number of ratings increase, the number of parameters and the number of sample points needed to estimate those parameters also increase. However, the recommendations in this paper should still apply.

Exhibits

Table 1

Simulation Results When Using Local Linear Regression, Including Other Rating as a Covariate, Data-Generating Model 1

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Mean Estimate of $Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.406	0.006	0.016	0.011	1.000	0.016	1.000
		Fuzzy IV	0.402	0.002	0.060	0.057	3.729	0.060	3.722
		Binding	0.403	0.003	0.015	0.009	0.959	0.015	0.957
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.403	0.003	0.015	0.011	1.000	0.015	1.000
		Fuzzy IV	0.396	-0.004	0.078	0.053	5.233	0.078	5.230
		Binding	0.403	0.003	0.015	0.009	1.038	0.015	1.038
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.399	-0.001	0.024	0.013	1.000	0.024	1.000
		Fuzzy IV	0.397	-0.003	0.071	0.064	2.966	0.071	2.966
		Binding	0.398	-0.002	0.013	0.008	0.543	0.013	0.543
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.389	-0.011	0.019	0.012	1.000	0.019	1.000
		Fuzzy IV	0.395	-0.005	0.082	0.064	4.245	0.082	4.219
		Binding	0.398	-0.002	0.013	0.008	0.676	0.013	0.672
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.392	-0.008	0.030	0.021	1.000	0.030	1.000
		Fuzzy IV	0.399	-0.001	0.104	0.108	3.503	0.104	3.495
		Binding	0.401	0.001	0.016	0.008	0.546	0.016	0.545
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.397	-0.003	0.031	0.021	1.000	0.031	1.000
		Fuzzy IV	0.393	-0.007	0.114	0.104	3.630	0.114	3.631
		Binding	0.401	0.001	0.016	0.008	0.517	0.016	0.517
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.395	-0.005	0.013	0.010	1.000	0.013	1.000
		Fuzzy IV	0.404	0.004	0.056	0.045	4.399	0.056	4.391
		Binding	0.402	0.002	0.011	0.008	0.881	0.011	0.879
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.409	0.009	0.019	0.012	1.000	0.019	1.000
		Fuzzy IV	0.397	-0.003	0.047	0.042	2.544	0.047	2.535
		Binding	0.402	0.002	0.011	0.008	0.601	0.011	0.599
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.398	-0.002	0.029	0.022	1.000	0.029	1.000
		Fuzzy IV	0.413	0.013	0.820	0.388	28.565	0.820	28.567
		Binding	0.396	-0.004	0.013	0.010	0.444	0.013	0.444
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.396	-0.004	0.016	0.011	1.000	0.016	1.000
		Fuzzy IV	0.385	-0.015	0.040	0.033	2.513	0.041	2.524
		Binding	0.396	-0.004	0.013	0.010	0.791	0.013	0.791

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a] Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table 2

Simulation Results When Using Local Linear Regression, Including Other Rating as a Covariate, Data-Generating Model 2

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Mean Estimate of $Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi_{R1}^{ATF} = 0.400$	Frontier	0.396	-0.004	0.025	0.022	1.000	0.025	1.000
		Fuzzy IV	0.328	-0.072 **	0.198	0.194	7.999	0.204	8.204
		Binding	0.389	-0.011	0.021	0.021	0.847	0.021	0.850
	$\psi_{R2}^{ATF} = 0.400$	Frontier	0.411	0.011	0.031	0.027	1.000	0.031	1.000
		Fuzzy IV	0.282	-0.118 **	0.725	0.444	23.242	0.739	23.600
		Binding	0.389	-0.011	0.021	0.021	0.672	0.021	0.674
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi_{R1}^{ATF} = 0.400$	Frontier	0.417	0.017 *	0.024	0.023	1.000	0.024	1.000
		Fuzzy IV	0.337	-0.063 **	0.157	0.150	6.560	0.161	6.646
		Binding	0.382	-0.018 **	0.022	0.019	0.935	0.023	0.936
	$\psi_{R2}^{ATF} = 0.400$	Frontier	0.396	-0.004	0.030	0.026	1.000	0.030	1.000
		Fuzzy IV	0.291	-0.109 **	0.419	0.364	13.854	0.431	14.243
		Binding	0.382	-0.018 **	0.022	0.019	0.738	0.023	0.748
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi_{R1}^{ATF} = 0.400$	Frontier	0.406	0.006	0.033	0.027	1.000	0.033	1.000
		Fuzzy IV	0.366	-0.034 *	0.111	0.121	3.339	0.112	3.369
		Binding	0.382	-0.018 **	0.022	0.019	0.671	0.023	0.680
	$\psi_{R2}^{ATF} = 0.400$	Frontier	0.424	0.024 **	0.028	0.024	1.000	0.029	1.000
		Fuzzy IV	0.380	-0.020	0.189	0.143	6.655	0.189	6.536
		Binding	0.382	-0.018 **	0.022	0.019	0.786	0.023	0.781
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi_{R1}^{ATF} = 0.400$	Frontier	0.413	0.013	0.023	0.022	1.000	0.023	1.000
		Fuzzy IV	0.374	-0.026	0.141	0.118	6.062	0.141	6.048
		Binding	0.408	0.008	0.023	0.020	1.008	0.023	1.003
	$\psi_{R2}^{ATF} = 0.400$	Frontier	0.409	0.009	0.037	0.032	1.000	0.037	1.000
		Fuzzy IV	0.339	-0.061 *	0.480	0.362	12.931	0.483	13.005
		Binding	0.408	0.008	0.023	0.020	0.631	0.023	0.631
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi_{R1}^{ATF} = 0.400$	Frontier	0.417	0.017	0.047	0.040	1.000	0.047	1.000
		Fuzzy IV	0.299	-0.101	1.460	1.254	31.377	1.471	31.393
		Binding	0.399	-0.001	0.032	0.033	0.689	0.032	0.685
	$\psi_{R2}^{ATF} = 0.400$	Frontier	0.370	-0.030 **	0.033	0.030	1.000	0.034	1.000
		Fuzzy IV	0.368	-0.032	0.301	0.258	9.159	0.302	8.952
		Binding	0.399	-0.001	0.032	0.033	0.977	0.032	0.952

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a]Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table 3

Simulation Results When Using Local Linear Regression, Including Other Rating as a Covariate, Data-Generating Model 3

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Mean Estimate of $Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.244$	Frontier	0.245	0.001	0.016	0.011	1.000	0.016	1.000
		Fuzzy IV	0.254	0.011	0.061	0.052	3.695	0.061	3.701
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.322$	Frontier	0.332	0.010	0.016	0.012	1.000	0.017	1.000
		Fuzzy IV	0.324	0.002	0.066	0.056	4.039	0.066	4.016
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.262$	Frontier	0.264	0.002	0.020	0.020	1.000	0.020	1.000
		Fuzzy IV	0.258	-0.004	0.071	0.071	3.497	0.071	3.497
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.331$	Frontier	0.329	-0.002	0.017	0.017	1.000	0.017	1.000
		Fuzzy IV	0.345	0.014	0.077	0.077	4.508	0.077	4.519
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.330$	Frontier	0.325	-0.006	0.026	0.020	1.000	0.026	1.000
		Fuzzy IV	0.319	-0.011	0.096	0.100	3.672	0.096	3.672
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.365$	Frontier	0.372	0.007	0.027	0.020	1.000	0.027	1.000
		Fuzzy IV	0.357	-0.009	0.107	0.108	3.948	0.107	3.943
		Binding	NA	NA	NA	NA	NA	NA	NA
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.209$	Frontier	0.206	-0.003	0.017	0.010	1.000	0.017	1.000
		Fuzzy IV	0.205	-0.004	0.064	0.044	3.704	0.064	3.703
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.304$	Frontier	0.305	0.001	0.018	0.011	1.000	0.018	1.000
		Fuzzy IV	0.295	-0.010	0.046	0.039	2.584	0.047	2.590
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.281$	Frontier	0.274	-0.007	0.028	0.022	1.000	0.028	1.000
		Fuzzy IV	0.341	0.059	1.442	1.160	51.472	1.446	51.496
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.294$	Frontier	0.294	0.000	0.017	0.010	1.000	0.017	1.000
		Fuzzy IV	0.297	0.003	0.050	0.036	2.865	0.050	2.866
		Binding	NA	NA	NA	NA	NA	NA	NA

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a] Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table 4

Simulation Results When Using Local Linear Regression, Including Other Rating as a Covariate, Data-Generating Model 4

Cuts/ρ	Parameter	Method	E($\hat{\psi}$)	Bias	Var($\hat{\psi}$)	Mean Estimate of Var($\hat{\psi}$)	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.167$	Frontier	0.167	0.000	0.025	0.024	1.000	0.025	1.000
		Fuzzy IV	0.110	-0.057 **	0.177	0.175	7.001	0.180	7.128
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.245$	Frontier	0.270	0.025 **	0.025	0.025	1.000	0.026	1.000
		Fuzzy IV	0.183	-0.062	0.576	0.476	22.717	0.580	22.333
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.202$	Frontier	0.218	0.016 *	0.030	0.023	1.000	0.030	1.000
		Fuzzy IV	0.137	-0.065 **	0.182	0.148	6.125	0.186	6.213
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.271$	Frontier	0.281	0.010	0.030	0.025	1.000	0.030	1.000
		Fuzzy IV	0.224	-0.047	0.537	0.379	17.965	0.539	17.972
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.315$	Frontier	0.336	0.021 *	0.037	0.027	1.000	0.037	1.000
		Fuzzy IV	0.317	0.002	0.113	0.116	3.098	0.113	3.060
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.350$	Frontier	0.355	0.005	0.035	0.026	1.000	0.035	1.000
		Fuzzy IV	0.308	-0.042 *	0.154	0.142	4.428	0.156	4.476
		Binding	NA	NA	NA	NA	NA	NA	NA
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.101$	Frontier	0.115	0.015 *	0.025	0.021	1.000	0.025	1.000
		Fuzzy IV	0.366	0.267 **	0.111	0.121	4.408	0.182	7.162
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.196$	Frontier	0.233	0.037 **	0.036	0.029	1.000	0.037	1.000
		Fuzzy IV	0.380	0.184 **	0.189	0.143	5.276	0.223	5.986
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.235$	Frontier	0.260	0.025 *	0.056	0.043	1.000	0.057	1.000
		Fuzzy IV	0.366	0.131 **	0.111	0.121	1.980	0.128	2.263
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.164$	Frontier	0.146	-0.017 *	0.036	0.030	1.000	0.036	1.000
		Fuzzy IV	0.380	0.216 **	0.189	0.143	5.263	0.235	6.505
		Binding	NA	NA	NA	NA	NA	NA	NA

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a] Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table 5

Application Study Results

Method	Parameter	$\%bw_l$	$\%bw_r$	N_{bw}	ψ	$\hat{se}_{OLS}$	$\hat{se}_{BS}$	$\mathrm{C}I_{OLS}$	$\mathrm{C}I_{BS}$
Frontier	$\psi_{CST\ ELA}^{ATF}$	0.4	0.5	2115	-2.20	2.20	4.40	[-6.6,2.2]	[-11.2,4.8]
Frontier	$\psi_{CELDT\ reading}^{ATF}$	0.6	0.9	2788	-1.90	3.10	3.60	[-7.9,2.5]	[-9.3,4.7]
Binding score	ψ_{BS}^{ATF}	0.4	0.6	3004	0.40	2.10	2.70	[-3.8,4.5]	[-4.0,6.4]

Figure 1

Design of Interest for 2RRDD

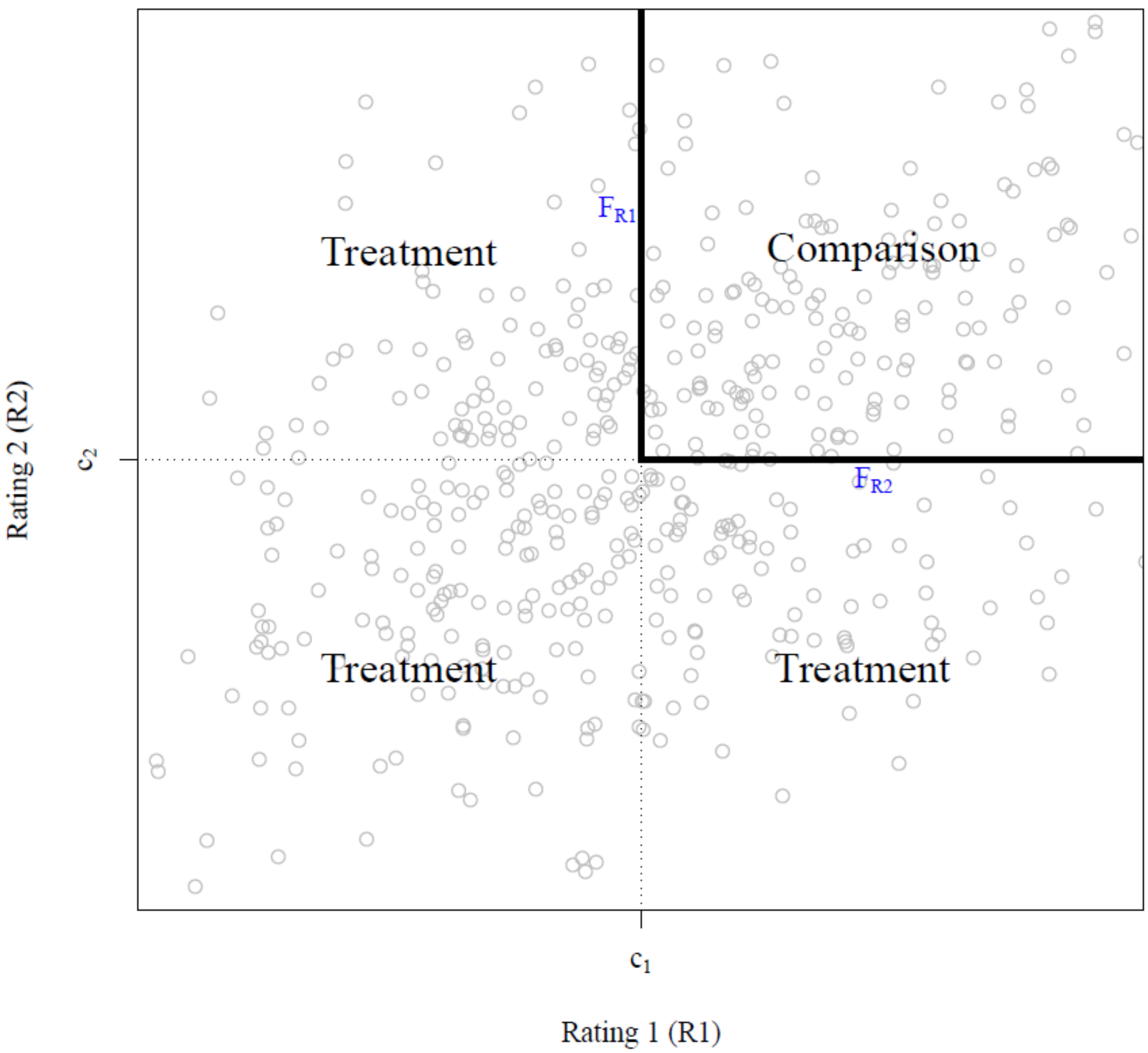
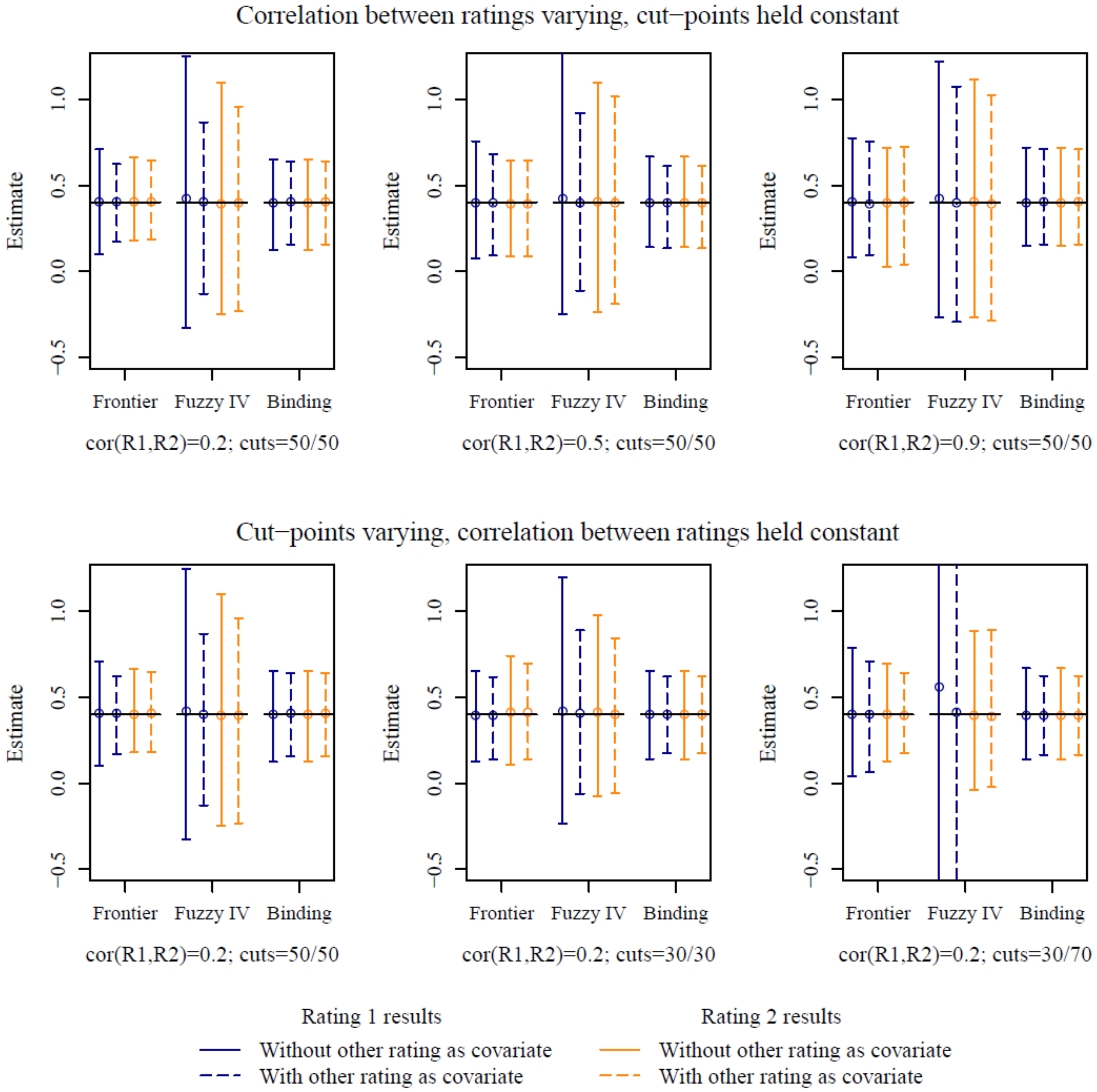


Figure 2

Simulation Results When Using Local Linear Regression, Data-Generating Model 1


Correlation between ratings varying, cut−points held constant

cor(R1,R2)=0.2; cuts=50/50

cor(R1,R2)=0.5; cuts=50/50

cor(R1,R2)=0.9; cuts=50/50

Cut−points varying, correlation between ratings held constant

cor(R1,R2)=0.2; cuts=50/50

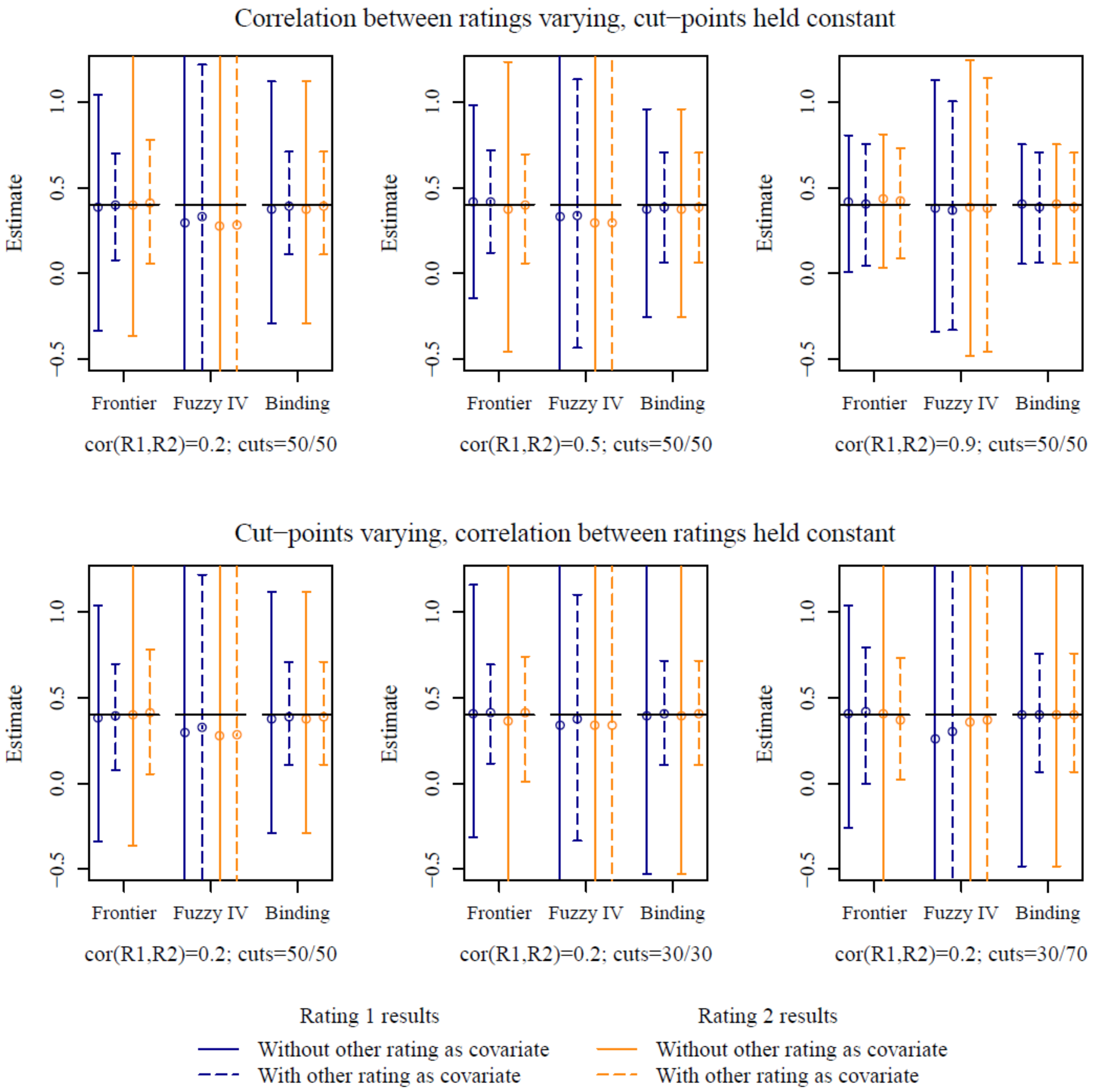
cor(R1,R2)=0.2; cuts=30/30

cor(R1,R2)=0.2; cuts=30/70

NOTE: For 500 samples of size 5,000.

Figure 3

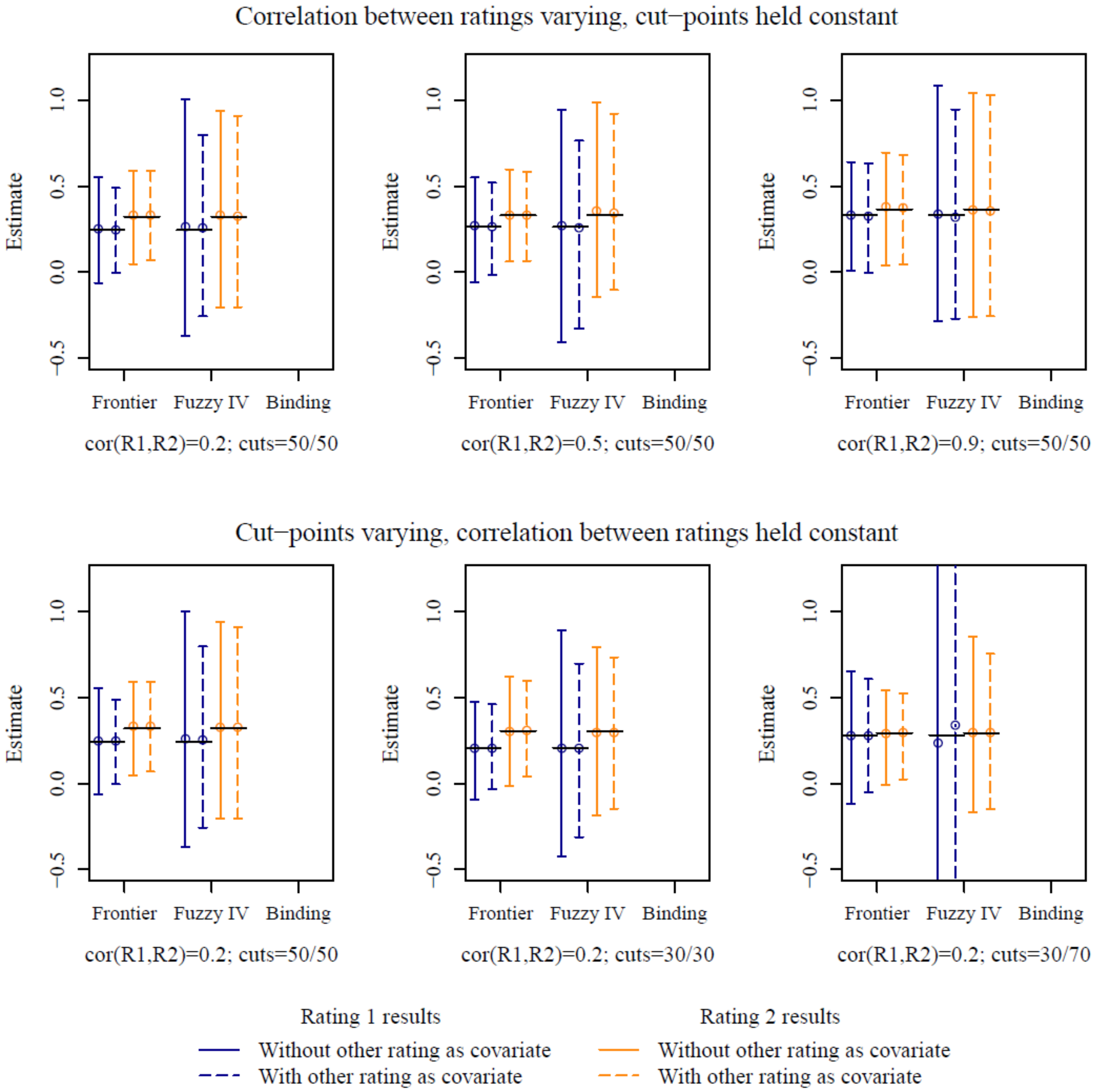
Simulation Results When Using Local Linear Regression, Data-Generating Model 2



NOTE: For 500 samples of size 5,000.

Figure 4

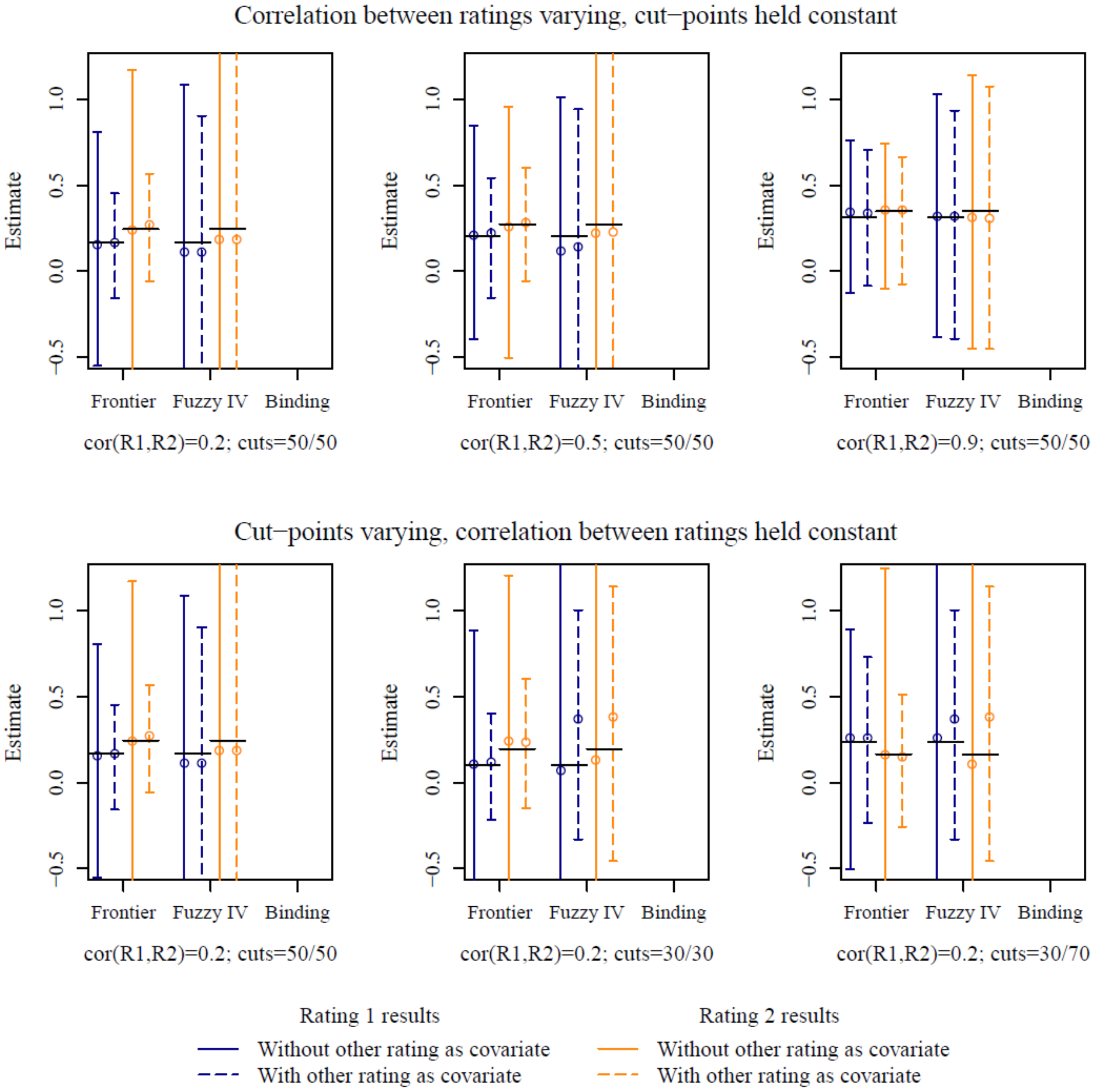
Simulation Results When Using Local Linear Regression, Data-Generating Model 3



NOTE: For 500 samples of size 5,000.

Figure 5

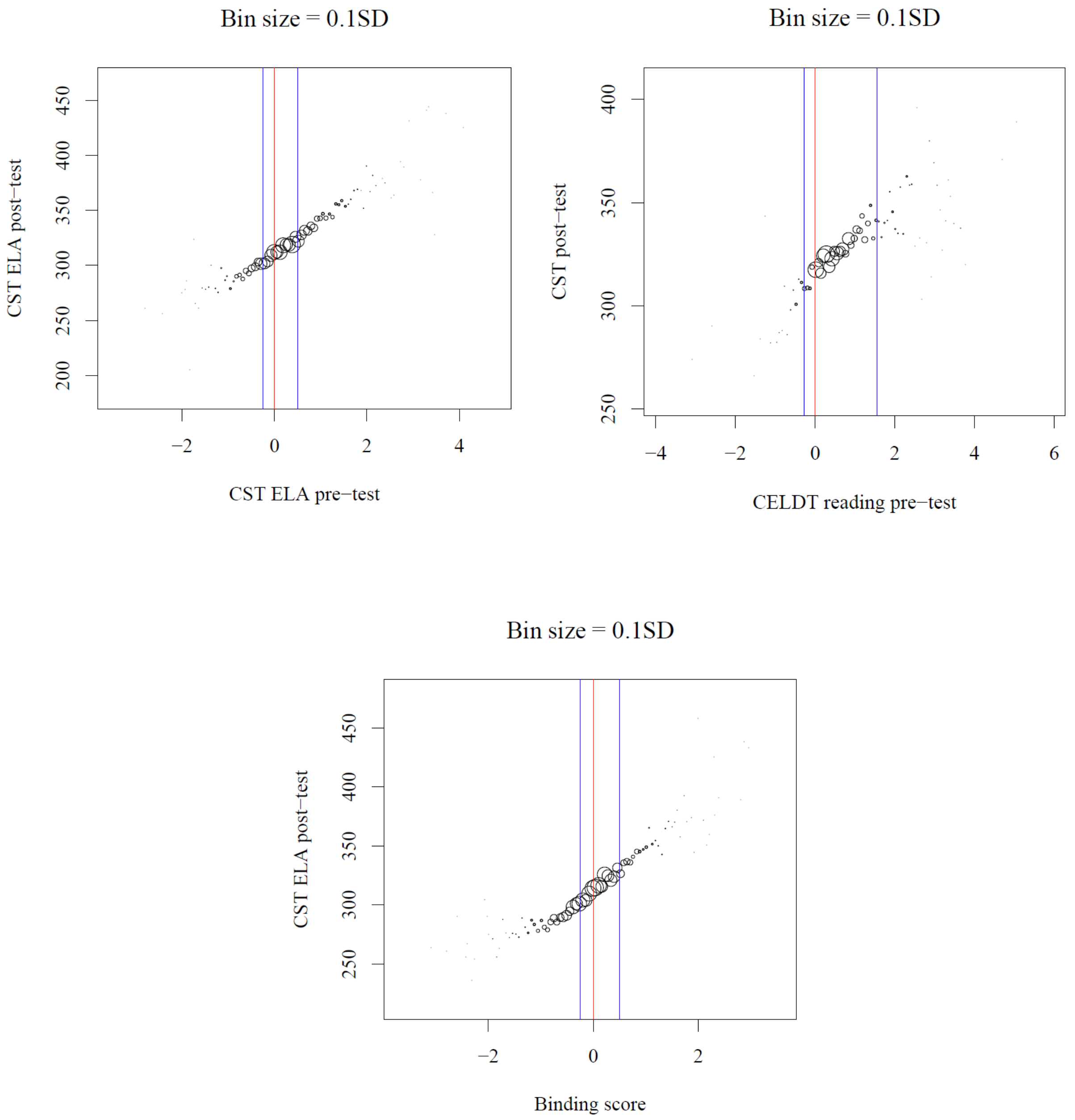
Simulation Results When Using Local Linear Regression, Data-Generating Model 4



NOTE: For 500 samples of size 5,000.

Figure 6

Relationship Between Outcome (CST ELA Post-Test) and Ratings



Bin size = 0.1SD

Bin size = 0.1SD

Bin size = 0.1SD

Appendix A

Computing Average Frontier Effects for Models 3 and 4 in Our Simulations

For Models 3 and 4 we used numerical integration to compute ψ_{R1}^{ATF} as a density-weighted mean of the conditional treatment effects for the values of rating 2 along the frontier (as in Wong, Steiner, and Cook, 2013), and we used the corresponding numerical integration to compute ψ_{R2}^{ATF} for the frontier for rating 2. In symbols:

$$\psi_{R1}^{ATF} = \frac{\int_{c2}^{\infty} g(c1,R2)\rho(c1,R2)dR2}{\int_{c2}^{\infty} \rho(c1,R2)dR2}$$

and

$$\psi_{R2}^{ATF} = \frac{\int_{c1}^{\infty} g(R1,c2)\rho(R1,c2)dR1}{\int_{c1}^{\infty} \rho(R1,c2)dR1}, \tag{A.1}$$

where $g(R1, R2)$ is the conditional treatment-effect function (i.e., $g(R1, R2) = E(Y_1 - Y_0|R1, R2)$), and $\rho(R1, R2)$ is the joint density of $R1$ and $R2$. Consequently values for ψ_{R1}^{ATF} and ψ_{R2}^{ATF} vary across simulated scenarios.

Appendix B

Estimation in Theory: Using Full Information

B.1 Implementation of the Four Methods Using Full Information

These analyses used the known functional form of the model that generated data for each simulated sample and data for all simulated sample members to estimate parameters of interest.

The surface method. For the surface method we used all of the sample points for each simulated sample to fit a regression with the functional form of the corresponding data-generating model such that:

For Model 1: $$Y_i = \beta_1 T_i + \beta_2 R1_i + \beta_3 R2_i + \epsilon_i. \quad \text{(B.1)}$$

For Model 2: $$Y_i = \beta_1 T_i + \beta_2 R1_i + \beta_3 R2_i + \beta_4 R1_i^2 + \beta_5 R2_i^2 + \epsilon_i\,. \quad \text{(B.2)}$$

For Model 3: $$Y_i = \beta_1 T_i + \beta_2 R1_i + \beta_3 R2_i + \beta_4 T_i R1_i + \beta_5 T_i R2_i + \epsilon_i. \quad \text{(B.3)}$$

For Model 4: $$\begin{aligned} Y_i = {} & \beta_1 T_i + \beta_2 R1_i + \beta_3 R2_i + \beta_4 R1_i^2 + \beta_5 R2_i^2 + \beta_6 T_i R1_i \\ & + \beta_7 T_i R2_i + \beta_8 T_i R1_i^2 + \beta_9 T_i R2_i^2 + \epsilon_i. \end{aligned} \quad \text{(B.4)}$$

In Models 1 and 2, the estimated treatment effects are equal to the estimated coefficient on the treatment status variable. In Models 3 and 4, however, the treatment effects depend on the values of $R1$ and $R2$ and differ, on average, at the two frontiers. We computed estimated average treatment effects for these models using numeric integration as described in Appendix A.[1]

The frontier method. The first step in the frontier method was to partition sample members by their status on one of the two ratings. Thus to estimate ψ_{R1}^{ATF} we only used observations for which $R2 \geq c2$, and to estimate ψ_{R2}^{ATF} we only used observations for which $R1 \geq c1$. We then fit the regression models that were used for the surface method, matching the functional form of the regression model to the known functional form of the data-generating model. Note that all of these models include both ratings.[2] For Models 3 and 4, which specify heterogeneous treatment effects that vary as a function of $R1$ and $R2$, we again used numerical integration after fitting our regression models, in order to obtain estimates of ψ_{R1}^{ATF} and ψ_{R2}^{ATF}.

The fuzzy IV method. The fuzzy IV method used all data points for each simulated sample. We defined an instrumental variable for treatment assignment $(1 - z1)$ based on rating

[1] In these calculations, we used knowledge of the shape of the rating distribution (bivariate normality) and estimated the conditional mean and conditional standard deviation of each rating from the simulated sample data. Because the surface models are fit using the functional form that generated the data that were simulated, these estimates used full information about functional forms.

[2] In theory, given the known bivariate distribution of the two correlated ratings, it is also possible to derive the functional form for a given frontier that depends only on the rating for that frontier.

1 such that $z1 \equiv I(R1 < 0)$. The correct specifications for the fuzzy IV models are provided in Appendix C.

The binding-score method. Recall that the binding-score method only produces results that can be compared with those produced by the other methods when treatment effects are homogeneous (as in Models 1 and 2). In order to assess the precision of the binding score method under ideal conditions (when the true functional form is known to the analyst) we need to express the outcome Y as a known function of terms that involve the binding score. While in theory it is possible to derive the true functional form of a binding-score model under full information, it will generally not be expressible as an easily parameterized function of terms that involve the binding score. For this reason, we did not include the binding-score method in this part of our simulation study.

B.2 Simulation Results Under Full Information

Figures B.1 through B.4 present simulation results for analyses that use our knowledge of the underlying data-generating models. Figure B.1 presents results for simulations under data-generating Model 1, which is linear in $R1$ and $R2$ and has a constant treatment effect. The top row in the figure reports distributions of estimates of ψ_{R1}^{ATF} (the darker lines) and ψ_{R2}^{ATF} (the lighter lines) when varying the correlation between $R1$ and $R2$ (0.2, 0.5, and 0.9) while holding the $R1$ and $R2$ cut-points constant at their median value. The bottom row presents distributions of estimates of ψ_{R1}^{ATF} and ψ_{R2}^{ATF} when varying the locations of the cut-points while holding the correlation between $R1$ and $R2$ at 0.20.

The bottom and top of the vertical lines for each estimation method and parameter (at the rating 1 and rating 2 frontiers) represent the 2.5th and 97.5th percentile values of the distribution of mean program effect estimates ($\hat{\psi}_{R1}^{ATF}$ and $\hat{\psi}_{R2}^{ATF}$) for 500 simulated samples. The point near the middle of each vertical line is the mean value of the 500 sample estimates. This value of $E(\hat{\psi}_{R1}^{ATF})$ or $E(\hat{\psi}_{R2}^{ATF})$ can be compared with the true value of the corresponding parameter, which is indicated by the solid horizontal line (at 0.4 for all simulations under Model 1). The distance of $E(\hat{\psi}_{R1}^{ATF})$ or $E(\hat{\psi}_{R2}^{ATF})$ from the horizontal line indicates the estimated bias for each estimator. The length of the line around these points indicates the precision of each estimator. Corresponding findings for data-generating Models 2, 3, and 4 are reported in Figures B.2 through B.4. More detailed findings for all simulations under full information are reported in Tables B.1 through B.4.

The first thing to note about these findings is that all three estimation methods produced unbiased estimates under all simulated scenarios, with only a few *apparent* exceptions. This is expected given our knowledge of the correct functional forms of these models. For the surface and frontier methods, we specify the correct functional form of the surface response model and the correct functional form of the rating distribution. For the fuzzy IV method, we specify the correct functional form of the first-stage model, the correct functional form of the reduced-form model, and the correct functional form of the rating distribution. Overall, as one looks across the full set of findings, 6 out of 120 tests for bias are statistically significant at the $p < 0.05$ level.[3] This is exactly the proportion (1 in 20) that should occur by chance if there were no bias. Furthermore, when we regenerated the data with a different seed, these cases no longer demonstrated statistically significant "bias." Thus, there is no evidence of systematic bias in any of the estimators under any of the data-generating scenarios.

In terms of precision, the figures indicate that the surface method is always superior to the others that were examined. It is more precise than the frontier method because the frontier method uses only part of a given sample whereas the surface method uses the whole sample. The surface method is more precise than the fuzzy IV method because of the added uncertainty produced by the two-stage IV estimation process. Indeed the fuzzy IV method was always the least precise of the methods that were examined. Thus even though it uses all sample points and therefore could theoretically be more precise than the frontier method (Reardon and Robinson, 2012), this was never the case in our simulation analyses.

[3] To test whether the estimated bias was statistically significant, we conducted a t-test across all 500 sample estimates.

Table B.1

Simulation Results with Full Information, Data-Generating Model 1

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant								
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Surface	0.402	0.002	0.007	1.000	0.007	1.000
		Frontier	0.399	-0.001	0.021	2.970	0.021	2.968
		Fuzzy IV	0.399	-0.001	0.041	5.901	0.041	5.898
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.402	0.002	0.007	1.000	0.007	1.000
		Frontier	0.398	-0.002	0.018	2.598	0.018	2.597
		Fuzzy IV	0.388	-0.012	0.038	5.490	0.038	5.507
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Surface	0.398	-0.002	0.009	1.000	0.009	1.000
		Frontier	0.401	0.001	0.022	2.345	0.022	2.344
		Fuzzy IV	0.409	0.009	0.047	4.979	0.047	4.985
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.398	-0.002	0.009	1.000	0.009	1.000
		Frontier	0.393	-0.007	0.021	2.248	0.021	2.253
		Fuzzy IV	0.393	-0.007	0.045	4.835	0.046	4.838
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Surface	0.393	-0.007	0.012	1.000	0.012	1.000
		Frontier	0.383	-0.017 *	0.026	2.273	0.027	2.287
		Fuzzy IV	0.401	0.001	0.068	5.897	0.068	5.871
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.393	-0.007	0.012	1.000	0.012	1.000
		Frontier	0.398	-0.002	0.027	2.352	0.027	2.343
		Fuzzy IV	0.432	0.032 **	0.067	5.750	0.068	5.814
Cut-points varying, correlation between ratings held constant								
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.400$	Surface	0.402	0.002	0.009	1.000	0.009	1.000
		Frontier	0.398	-0.002	0.018	1.962	0.018	1.962
		Fuzzy IV	0.394	-0.006	0.031	3.474	0.031	3.475
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.402	0.002	0.009	1.000	0.009	1.000
		Frontier	0.409	0.009	0.018	1.972	0.018	1.979
		Fuzzy IV	0.411	0.011	0.028	3.114	0.028	3.126
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.400$	Surface	0.394	-0.006	0.011	1.000	0.011	1.000
		Frontier	0.384	-0.016	0.042	3.940	0.043	3.948
		Fuzzy IV	0.378	-0.022	0.190	17.609	0.190	17.587
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.394	-0.006	0.011	1.000	0.011	1.000
		Frontier	0.400	0.000	0.015	1.387	0.015	1.381
		Fuzzy IV	0.400	0.000	0.023	2.110	0.023	2.102

Note: For 500 samples of size 1,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a]Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table B.2

Simulation Results with Full Information, Data-Generating Model 2

						Relative		Relative
Cuts/ρ	Parameter	Method	E($\hat{\psi}$)	Bias	Var($\hat{\psi}$)	Var[a]	MSE	MSE[a]
Correlation between ratings varying, cut-points held constant								
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Surface	0.399	-0.001	0.009	1.000	0.009	1.000
		Frontier	0.399	-0.001	0.021	2.212	0.021	2.212
		Fuzzy IV	0.376	-0.024	0.139	14.952	0.140	15.013
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.399	-0.001	0.009	1.000	0.009	1.000
		Frontier	0.391	-0.009	0.024	2.627	0.025	2.635
		Fuzzy IV	0.404	0.004	0.389	41.851	0.389	41.846
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Surface	0.401	0.001	0.008	1.000	0.008	1.000
		Frontier	0.397	-0.003	0.023	2.822	0.023	2.823
		Fuzzy IV	0.399	-0.001	0.130	15.674	0.130	15.673
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.401	0.001	0.008	1.000	0.008	1.000
		Frontier	0.393	-0.007	0.024	2.932	0.024	2.938
		Fuzzy IV	0.394	-0.006	0.368	44.461	0.368	44.464
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Surface	0.401	0.001	0.012	1.000	0.012	1.000
		Frontier	0.399	-0.001	0.047	3.931	0.047	3.930
		Fuzzy IV	0.379	-0.021	0.095	8.002	0.096	8.039
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.401	0.001	0.012	1.000	0.012	1.000
		Frontier	0.407	0.007	0.045	3.746	0.045	3.750
		Fuzzy IV	0.411	0.011	0.175	14.715	0.176	14.725
Cut-points varying, correlation between ratings held constant								
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.400$	Surface	0.403	0.003	0.010	1.000	0.010	1.000
		Frontier	0.404	0.004	0.022	2.258	0.022	2.258
		Fuzzy IV	0.393	-0.007	0.099	10.138	0.099	10.136
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.403	0.003	0.010	1.000	0.010	1.000
		Frontier	0.404	0.004	0.022	2.289	0.022	2.288
		Fuzzy IV	0.357	-0.043	0.316	32.458	0.318	32.619
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.400$	Surface	0.395	-0.005	0.013	1.000	0.013	1.000
		Frontier	0.396	-0.004	0.063	5.012	0.063	5.003
		Fuzzy IV	0.425	0.025	0.656	52.406	0.657	52.352
	$\psi^{ATF}_{R2} = 0.400$	Surface	0.395	-0.005	0.013	1.000	0.013	1.000
		Frontier	0.402	0.002	0.019	1.524	0.019	1.521
		Fuzzy IV	0.393	-0.007	0.216	17.226	0.216	17.196

Note: For 500 samples of size 1,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a]Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table B.3

Simulation Results with Full Information, Data-Generating Model 3

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant								
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.244$	Surface	0.253	0.009	0.012	1.000	0.013	1.000
		Frontier	0.248	0.005	0.022	1.756	0.022	1.745
		Fuzzy IV	0.264	0.020	0.095	7.657	0.096	7.635
	$\psi^{ATF}_{R2} = 0.322$	Surface	0.338	0.017 **	0.013	1.000	0.013	1.000
		Frontier	0.340	0.018 **	0.021	1.639	0.021	1.630
		Fuzzy IV	0.338	0.016	0.104	8.203	0.104	8.046
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.262$	Surface	0.259	-0.003	0.014	1.000	0.014	1.000
		Frontier	0.269	0.007	0.026	1.827	0.026	1.829
		Fuzzy IV	0.259	-0.003	0.118	8.389	0.118	8.385
	$\psi^{ATF}_{R2} = 0.331$	Surface	0.327	-0.004	0.015	1.000	0.015	1.000
		Frontier	0.319	-0.012	0.027	1.803	0.027	1.811
		Fuzzy IV	0.308	-0.023	0.115	7.783	0.115	7.810
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.330$	Surface	0.330	0.000	0.013	1.000	0.013	1.000
		Frontier	0.326	-0.004	0.042	3.116	0.042	3.117
		Fuzzy IV	0.336	0.005	0.166	12.367	0.166	12.369
	$\psi^{ATF}_{R2} = 0.356$	Surface	0.358	-0.007	0.012	1.000	0.012	1.000
		Frontier	0.353	-0.012	0.043	3.552	0.043	3.549
		Fuzzy IV	0.366	0.000	0.159	13.053	0.159	12.995
Cut-points varying, correlation between ratings held constant								
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.209$	Surface	0.202	-0.007	0.011	1.000	0.011	1.000
		Frontier	0.200	-0.009	0.023	2.019	0.023	2.017
		Fuzzy IV	0.202	-0.007	0.073	6.418	0.073	6.394
	$\psi^{ATF}_{R2} = 0.304$	Surface	0.302	-0.003	0.013	1.000	0.013	1.000
		Frontier	0.306	0.001	0.022	1.659	0.022	1.658
		Fuzzy IV	0.295	-0.010	0.074	5.613	0.075	5.617
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.281$	Surface	0.290	0.008	0.020	1.000	0.020	1.000
		Frontier	0.289	0.007	0.049	2.452	0.049	2.446
		Fuzzy IV	0.248	-0.033	0.473	23.613	0.474	23.586
	$\psi^{ATF}_{R2} = 0.294$	Surface	0.298	0.004	0.013	1.000	0.013	1.000
		Frontier	0.299	0.005	0.015	1.173	0.015	1.174
		Fuzzy IV	0.307	0.013	0.056	4.322	0.056	4.329

Note: For 500 samples of size 1,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a]Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table B.4

Simulation Results with Full Information, Data-Generating Model 4

						Relative		Relative
Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Var[a]	MSE	MSE[a]
		Correlation between ratings varying, cut-points held constant						
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.167$	Surface	0.182	0.015 *	0.020	1.000	0.021	1.000
		Frontier	0.182	0.015	0.037	1.822	0.038	1.812
		Fuzzy IV	0.183	0.016	0.497	24.273	0.497	24.012
	$\psi^{ATF}_{R2} = 0.245$	Surface	0.245	0.000	0.025	1.000	0.025	1.000
		Frontier	0.245	0.000	0.042	1.709	0.042	1.709
		Fuzzy IV	0.306	0.061	1.475	59.501	1.479	59.649
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.202$	Surface	0.191	-0.010	0.021	1.000	0.021	1.000
		Frontier	0.189	-0.012	0.048	2.306	0.048	2.302
		Fuzzy IV	0.168	-0.033	0.313	14.953	0.314	14.931
	$\psi^{ATF}_{R2} = 0.271$	Surface	0.272	0.001	0.023	1.000	0.023	1.000
		Frontier	0.275	0.004	0.049	2.115	0.049	2.116
		Fuzzy IV	0.237	-0.034	0.776	33.382	0.777	33.432
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.315$	Surface	0.324	0.009	0.025	1.000	0.025	1.000
		Frontier	0.339	0.024	0.084	3.299	0.084	3.310
		Fuzzy IV	0.324	0.009	0.249	9.829	0.249	9.800
	$\psi^{ATF}_{R2} = 0.350$	Surface	0.353	0.003	0.027	1.000	0.027	1.000
		Frontier	0.359	0.009	0.083	3.047	0.083	3.050
		Fuzzy IV	0.345	-0.005	0.450	16.523	0.450	16.519
		Cut-points varying, correlation between ratings held constant						
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.101$	Surface	0.106	0.005	0.018	1.000	0.018	1.000
		Frontier	0.100	-0.001	0.035	1.983	0.035	1.980
		Fuzzy IV	0.120	0.019	0.291	16.591	0.292	16.586
	$\psi^{ATF}_{R2} = 0.196$	Surface	0.196	0.000	0.018	1.000	0.018	1.000
		Frontier	0.198	0.002	0.040	2.212	0.040	2.212
		Fuzzy IV	0.232	0.035	0.900	49.922	0.901	49.991
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.235$	Surface	0.213	-0.022 *	0.041	1.000	0.041	1.000
		Frontier	0.222	-0.013	0.108	2.656	0.109	2.630
		Fuzzy IV	0.392	0.157 *	2.102	51.525	2.127	51.544
	$\psi^{ATF}_{R2} = 0.163$	Surface	0.170	0.006	0.026	1.000	0.026	1.000
		Frontier	0.160	-0.004	0.033	1.258	0.033	1.257
		Fuzzy IV	0.126	-0.038	0.730	28.089	0.731	28.110

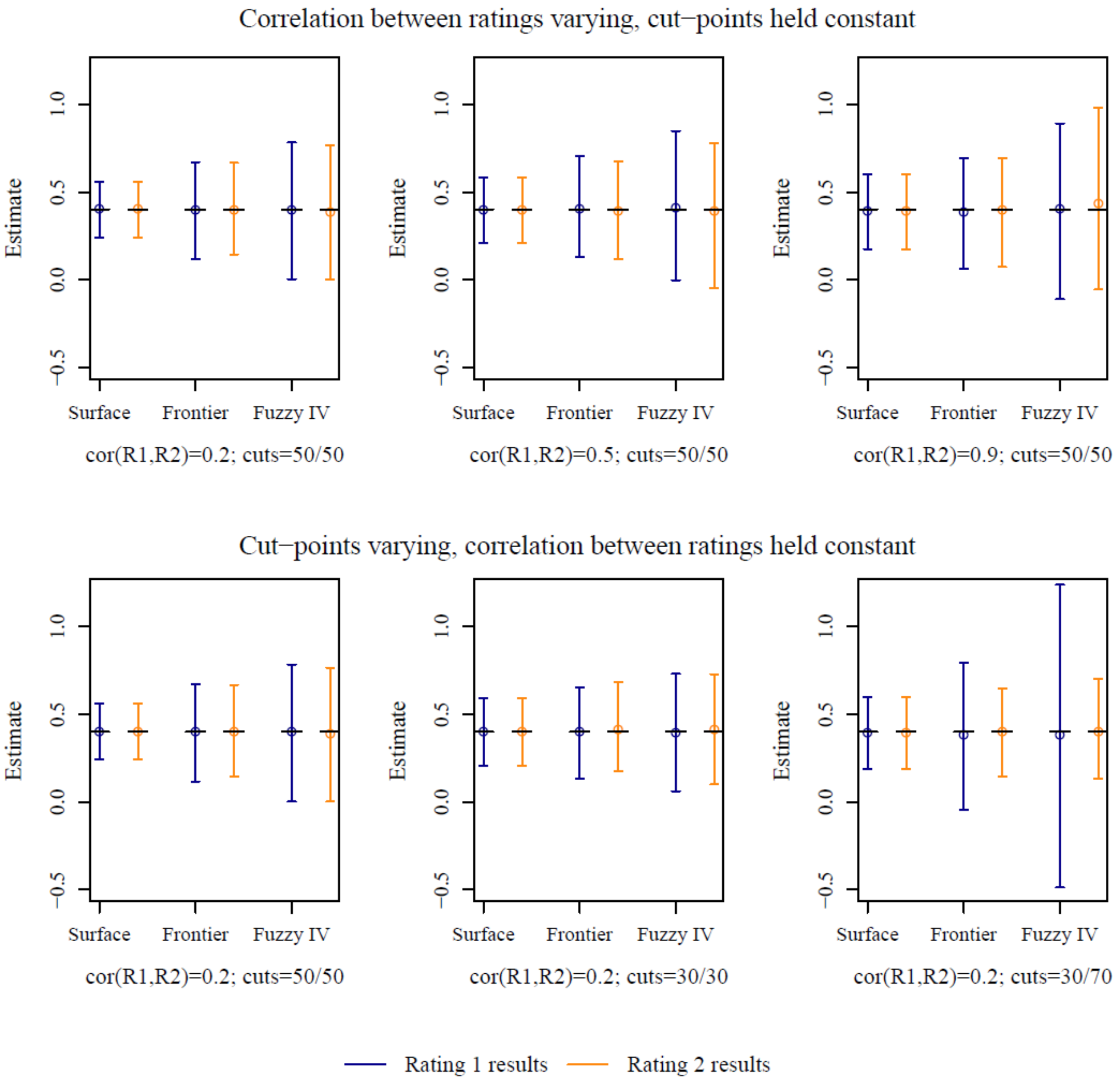
Note: For 500 samples of size 1,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a] Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Figure B.1

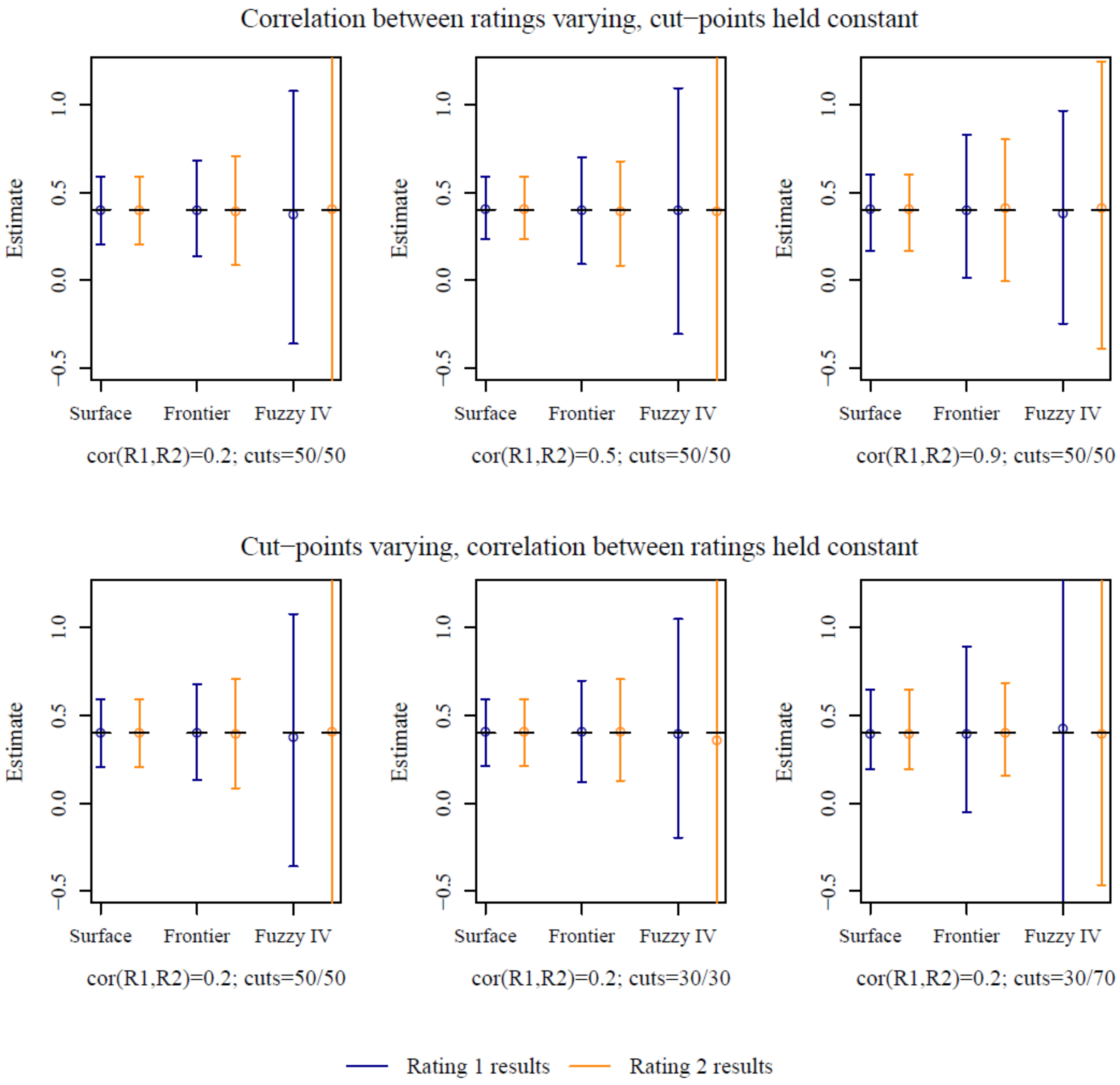
Simulation Results with Full Information, Data-Generating Model 1



NOTE: For 500 samples of size 1,000.

Figure B.2

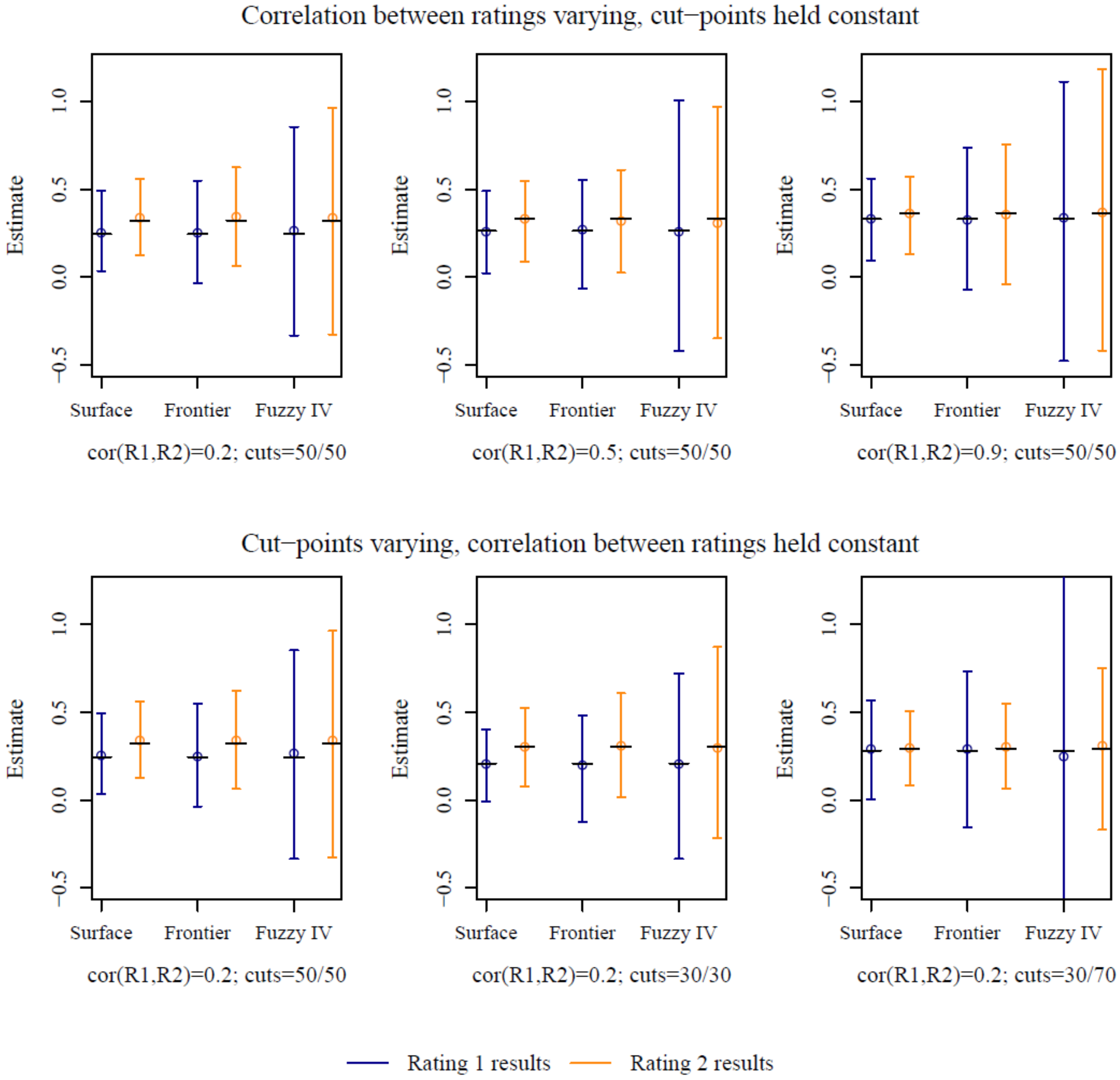
Simulation Results with Full Information, Data-Generating Model 2



NOTE: For 500 samples of size 1,000.

Figure B.3

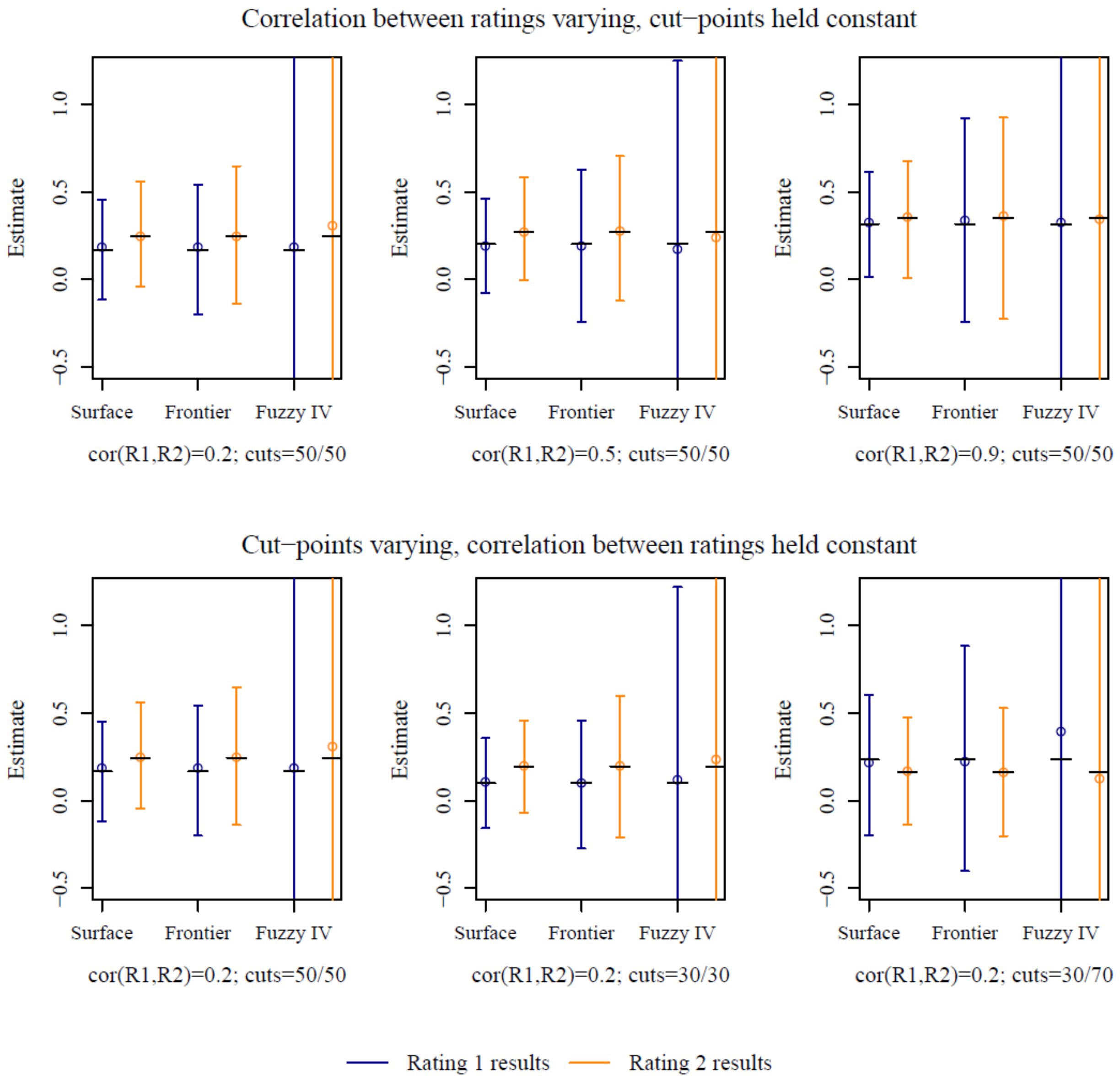
Simulation Results with Full Information, Data-Generating Model 3



NOTE: For 500 samples of size 1,000.

Figure B.4

Simulation Results with Full Information, Data-Generating Model 4



NOTE: For 500 samples of size 1,000.

Appendix C

Deriving Correct Model Specifications for the Fuzzy IV Method

This appendix derives correct specifications for the first-stage, second-stage, and reduced-form models used for fuzzy IV estimators in our analysis under full information.

Recall that we have two ratings, $R1$ and $R2$, that were generated from a bivariate normal distribution, with mean 0, variance 1, and a correlation $\rho_{R1,R2}$ (or simply ρ for short) that varies by simulation with values equal to 0.2, 0.5, or 0.9. After defining cut-points (as percentiles in the ratings) $(c1_p, c2_p)$, we centered the ratings around the cut-points. Therefore, the means of the ratings $R1$ and $R2$ and the actual cut-point values $(c1, c2)$ vary by simulation. We can then write the joint distribution of $R1$ and $R2$ as

$$\begin{pmatrix} R1 \\ R2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

Next, we define $z1 = I(R1 < 0)$ and $z2 = I(R2 < 0)$. Then, given our design of interest, subjects are assigned to treatment if $z1 = 1$ or $z2 = 1$. Equivalently, treatment indicator $T = \max(z1, z2)$.

For our derivations, we focus on the case in which we estimate ψ_{R1}^{ATF}, the treatment effect at the rating 1 frontier. The model specifications for estimating ψ_{R2}^{ATF} can be derived the same way.

To compute the Wald estimator for a fuzzy RDD design, we need to estimate both the discontinuity in the probability of treatment receipt at the cut-point and the discontinuity in the average value of the outcome at the cut-point. We estimate both using RDD models, and then take their ratio.

Estimating the Discontinuity in Treatment Probability

Given that the two ratings have a bivariate normal distribution, the proportion of units assigned to the treatment condition ($T = 1$) is a function of $R1$:

$$\begin{aligned} E(T|R1) &= 1 - (1 - z1)P(T = 0|R1) \\ &= 1 - (1 - z1)P(R2 \geq 0|R1) \\ &= 1 - (1 - z1)\Phi\left(\frac{\mu_2 + \rho(R1 - \mu_1)}{\sqrt{(1 - \rho^2)}}\right), \end{aligned} \tag{C.1}$$

where Φ is the standard normal cumulative distribution function.

Next let

$$f(R1) \equiv P(R2 \geq 0|R1) = \Phi\left(\frac{\mu_2+\rho(R1-\mu_1)}{\sqrt{(1-\rho^2)}}\right), \tag{C.2}$$

and

$$f(0) \equiv P(R2 \geq 0|R1 = 0) = \Phi\left(\frac{\mu_2+\rho(-\mu_1)}{\sqrt{(1-\rho^2)}}\right). \tag{C.3}$$

Therefore, $f(0)$ is the increase in the probability of receiving treatment as we cross the rating 1 threshold from the $z1 = 0$ region to the $z1 = 1$ region. We can now write (C.1) as follows:

$$\begin{aligned} E(T|R1) &= 1 - (1 - z1)f(R1) \\ &= 1 - (1 - z1)[f(0) + f(R1) - f(0)] \\ &= 1 - (1 - z1)f(0) + (1 - z1)[f(R1) - f(0)] \\ &= 1 - (1 - z1)f(0) - (1 - z1)[f(R1) - f(0)]. \end{aligned} \tag{C.4}$$

We write $E(T|R1)$ this way because we want to be able to estimate the parameter $f(0)$, the discontinuity in the probability of treatment assignment. Therefore, $(1 - z1)$ is the instrument. Next, let $R1' \equiv f(R1) - f(0)$. Then we can write the first-stage regression model as:

$$E(T|R1) = \gamma_0 + \gamma_1(1 - z1) + \gamma_2(1 - z1)R1', \tag{C.5}$$

such that $\gamma_1 = -f(0)$, $\gamma_0 = 1$, and $\gamma_2 = -1$. In practice, unless we were certain of the functional form, we would estimate the three parameters of this model. When we fit (C.5), $\hat{\gamma}_1$is an estimator of the jump in the probability of receiving treatment at the rating 1 threshold.

These relationships in the first-stage model are illustrated in the plot in Figure C.1. When $R1 < 0$, the probability of receiving treatment is 1 by design. In this case, in (C.5), $(1 - z1) = 0$, so $E(T|R1) = \gamma_0 = 1$. When $R1 \geq 0$, the probability of treatment depends on $P(R2 < 0|R1)$, by design. Now in (C.5), because $\gamma_0 = 1$ and $\gamma_2 = -1$, we have $E(T|R1) = 1 - \gamma_1 - R1' = 1 - \gamma_1 - [f(R1) - f(0)] = 1 - f(R1) = P(R2 < 0|R1)$. Given the correlation between $R1$ and $R2$, the probability of receiving treatment is greater for values of $R1$ closer to the rating 1 threshold. Finally, when $R1 = 0$, γ_1 (equivalent to $-f(0)$) is the length of the red line (the vertical segment at the cut-point between the intersections of the plotted curve). It is the discontinuity in the probability of receiving treatment at the rating 1 threshold.

Estimating the Discontinuity in Outcomes

Suppose the true data-generating surface response model is given by

Model 1: $$Y = \alpha_0 + \alpha_1 T + \alpha_2 R1 + \alpha_3 R2 + \epsilon. \tag{C.6}$$

Recall that with the fuzzy IV method, we treat the analysis as if we are in the setting of a single-rating RDD. Therefore, we want to model the conditional expectation of Y given $R1$, which is given by

$$\begin{aligned} E(Y|R1) &= \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 E(R1|R1) + \alpha_3 E(R2|R1) + E(\epsilon|R1) \\ &= \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 R1 + \alpha_3[\mu_2 + \rho(R1 - \mu_1)] \\ &= \delta_0 + \delta_1 E(T|R1) + \delta_2 R1. \end{aligned} \tag{C.7}$$

We can then plug the first-stage model into (C.7) so that we have the reduced-form equation given by

$$\begin{aligned} E(Y|R1) &= \delta_0 + \delta_1[1 + \gamma_1(1 - z1) + \gamma_2(1 - z1)R1'] + \delta_2 R1 \\ &= \beta_0 + \beta_1(1 - z1) + \beta_2(1 - z1)R1' + \beta_3 R1. \end{aligned} \tag{C.8}$$

When we fit (C.8), $\hat{\beta}_1$is an estimator of the difference in average outcomes at the rating 1 threshold. Therefore, the Wald estimator of ψ_{R1}^{ATF} is given by

$$\hat{\psi}_{R1}^{ATF}(Wald) = \frac{\hat{\beta}_1}{\hat{\gamma}_1}. \tag{C.9}$$

Note that $\hat{\psi}_{R1}^{ATF}(Wald)$ is a *consistent* estimator of ψ_{R1}^{ATF}, but is, like all IV estimators, subject to finite sample bias.

Next, suppose that the true data-generating surface response model is given by:

Model 2: $$Y = \alpha_0 + \alpha_1 T + \alpha_2 R1 + \alpha_3 R2 + \alpha_3 R1^2 + \alpha_3 R2^2 + \epsilon. \tag{C.10}$$

The conditional expectation of Y given $R1$ is given by

$$\begin{aligned} E(Y|R1) &= \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 E(R1|R1) + \alpha_3 E(R2|R1) + \alpha_4 E(R1^2|R1) \\ &\qquad + \alpha_5 E(R2^2|R1) + E(\epsilon|R1) \\ &= \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 R1 + \alpha_3[\mu_2 + \rho(R1 - \mu_1)] + \alpha_4 R1^2 + \alpha_5 E(R2^2|R1) \\ &= \delta_0 + \delta_1 E(T|R1) + \delta_2 R1 + \delta_3 R1^2 + \delta_4 h(R1), \end{aligned} \tag{C.11}$$

where

$$\begin{aligned} h(R1) \equiv E(R2^2|R1_i) &= \int_{-\infty}^{\infty} x^2 \Gamma(R1, x)dx \\ &= E(R2|R1)^2 + Var(R2|R1) \end{aligned}$$

$$
\begin{aligned}
&= \left(\mu_2 + \rho(R1 - \mu_1)\right)^2 + (1 - \rho^2) \\
&= (\mu_2 - \rho\mu_1)^2 + (1 - \rho^2) + 2\rho(\mu_2 - \rho\mu_1)R1 + \rho^2 R1^2.
\end{aligned}
\tag{C.12}
$$

Plugging the first-stage model (C.5) and (C.11) into (C.12), the reduced-form model is then given by

$$
\begin{aligned}
E(Y|R1) &= \delta_0 + \delta_1 E(T|R1) + \delta_2 R1 + \delta_3 R1^2 + \delta_4 h(R1) \\
&= \delta_0 + \delta_1[1 + \gamma_1(1 - z1) + \gamma_2(1 - z1)R1'] + \delta_2 R1 + \delta_3 R1^2 \\
&\qquad + \delta_4[(\mu_2 - \rho\mu_1)^2 + (1 - \rho^2) + 2\rho(\mu_2 - \rho\mu_1)R1 + \rho^2 R1^2] \\
&= \beta_0 + \beta_1(1 - z1) + \beta_2 R1 + \beta_3 R1^2 + \beta_4(1 - z1)R1'.
\end{aligned}
\tag{C.13}
$$

We then fit (C.13) and compute the Wald estimator as $\hat{\psi}_{R1}^{ATF}(Wald) = \frac{\widehat{\beta_1}}{\widehat{\gamma_1}}$.

Next, suppose that the true data-generating surface response model is instead given by

Model 3: $\quad Y = \alpha_0 + \alpha_1 T + \alpha_2 R1 + \alpha_3 R2 + \alpha_4 T \cdot R1 + \alpha_5 T \cdot R2 + \epsilon. \quad$ (C.14)

Now, the conditional expectation of Y given $R1_S$ is given by

$$
\begin{aligned}
E(Y|R1) &= \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 E(R1|R1) + \alpha_3 E(R2|R1) + \alpha_4 E(T \cdot R1|R1) \\
&\qquad + \alpha_5 E(T \cdot R2|R1) + E(\epsilon|R1) \\
&= \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 R1 + \alpha_3[\mu_2 + \rho(R1 - \mu_1)] + \alpha_4 R1 E(T|R1) \\
&\qquad + \alpha_5 E(T \cdot R2|R1) \\
&= \delta_0 + \delta_1 E(T|R1) + \delta_2 R1 + \delta_3 R1 E(T|R1) + \delta_4 g(R1),
\end{aligned}
\tag{C.15}
$$

where

$$
\begin{aligned}
g(R1) &\equiv E(T \cdot R2|R1) \\
&= z1 \cdot E(R2|R1) + (1 - z1) \cdot P(T = 1|R1) \cdot E(R2|R1, R2 < 0) \\
&= z1 \cdot E(R2|R1) + (1 - z1) \cdot \left(1 - f(R1)\right) \\
&\qquad \cdot \frac{1}{1 - f(R1)} \int_{-\infty}^{0} x \cdot \phi\left(\frac{\mu_2 + \rho(R1 - \mu_1) - x}{\sqrt{1 - \rho^2}}\right) dx \\
&= z1[\mu_2 + \rho(R1 - \mu_1)] + (1 - z1) \int_{-\infty}^{0} x \cdot \phi\left(\frac{\mu_2 + \rho(R1 - \mu_1) - x}{\sqrt{1 - \rho^2}}\right) dx \\
&= z1[\mu_2 + \rho(R1 - \mu_1)] + (1 - z1) g'(R1),
\end{aligned}
\tag{C.16}
$$

where $g'(R1) = \int_{-\infty}^{0} x \cdot \phi\left(\frac{\mu_2 + \rho(R1 - \mu_1) - x}{\sqrt{1 - \rho^2}}\right) dx$. To obtain the reduced-form equation, we plug (C.5) and (C.15) into (C.16), so we have

$$\begin{aligned}
E(Y|R1) = {} & \delta_0 + \delta_1[1 + \gamma_1(1 - z1) + \gamma_2(1 - z1)R1'] + \delta_2 R1 \\
& + \delta_3 R1[1 + \gamma_1(1 - z1) + \gamma_2(1 - z1)R1'] \\
& + \delta_4[z1[\mu_2 + \rho(R1 - \mu_1)] + (1 - z1)g'(R1)] \\
= {} & \beta_0 + \beta_1(1 - z1) + \beta_2 R1 + \beta_3(1 - z1)R1 + \beta_4(1 - z1)R1' + \beta_5(1 - z1)R1 \\
& \cdot R1' + \beta_6(1 - z1)g'(R1).
\end{aligned} \tag{C.17}$$

We then fit (C.17) and compute the Wald estimator as above: $\widehat{\psi}_{R1}^{ATF}(Wald) = \frac{\widehat{\beta_1}}{\widehat{\gamma_1}}$.

Finally, suppose that the true data-generating surface response model is instead given by

Model 4:
$$\begin{aligned}
Y = {} & \alpha_0 + \alpha_1 T + \alpha_2 R1 + \alpha_3 R2 + \alpha_4 R1^2 + \alpha_5 R2^2 + \alpha_6 TR1 + \\
& \alpha_7 TR2 + \alpha_8 TR1^2 + \alpha_9 TR2^2 + \epsilon.
\end{aligned} \tag{C.18}$$

In this case, the conditional expectation of Y given $R1$ can be written as

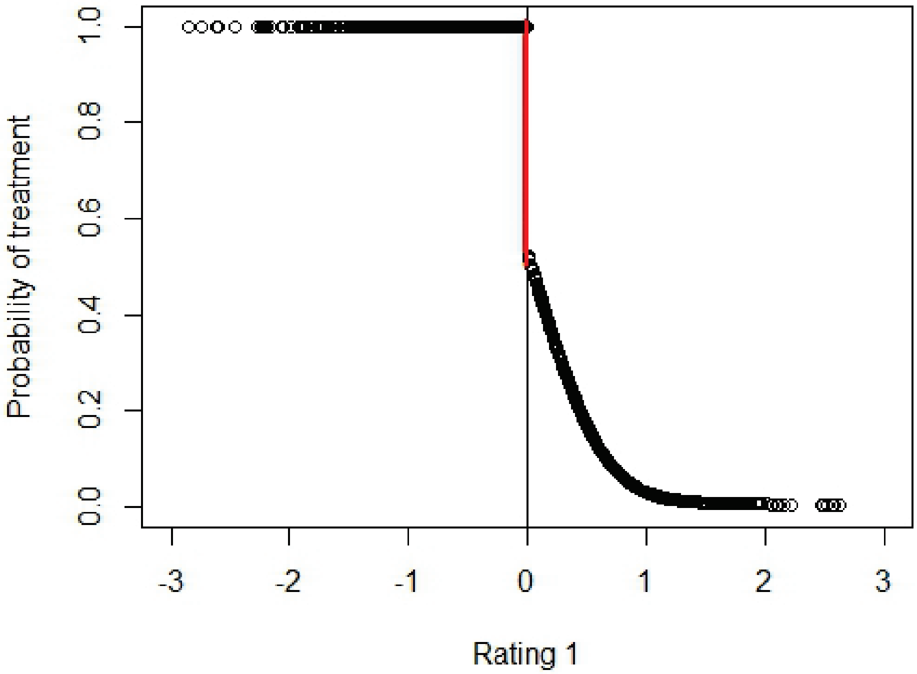
$$\begin{aligned}
E(Y|R1) = {} & \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 E(R1|R1) + \alpha_3 E(R2|R1) + \alpha_4 E(R1^2|R1) \\
& + \alpha_5 E(R2^2|R1) + \alpha_6 E(T \cdot R1|R1) + \alpha_7 E(T \cdot R2|R1) + \alpha_8 E(T \cdot R1^2|R1) \\
& + \alpha_9 E(T \cdot R2^2|R1) + E(\epsilon|R1) \\
= {} & \alpha_0 + \alpha_1 E(T|R1) + \alpha_2 R1 + \alpha_3[\mu_2 + \rho(R1 - \mu_1)] + \alpha_4 R1^2 + \alpha_5 E(R2^2|R1) \\
& + \alpha_6 R1 E(T|R1) + \alpha_7 E(T \cdot R2|R1) + \alpha_8 R1^2 E(T|R1) + \alpha_9 E(T \cdot R2^2|R1) \\
= {} & \delta_0 + \delta_1 E(T|R1) + \delta_2 R1 + \delta_3 R1^2 + \delta_4 E(R2^2|R1) + \delta_5 R1 E(T|R1) \\
& + \delta_6 g(R1) + \delta_7 R1^2 E(T|R1) + \delta_8 G(R1) \\
= {} & \delta_0 + (\delta_1 + \delta_5 R1 + \delta_7 R1^2)[1 + \gamma_1(1 - z1) + \gamma_2(1 - z1)R1'] + \delta_2 R1 \\
& + \delta_3 R1^2 + (\delta_4 + \delta_8 z1)\left[\left(\mu_2 + \rho(R1 - \mu_1)\right)^2 + (1 - \rho^2)\right] \\
& + \delta_6[z1[\mu_2 + \rho(R1 - \mu_1)] + (1 - z1)g'(R1)] + \delta_8(1 - z1)g''(R1) \\
= {} & \beta_0 + \beta_1(1 - z1) + \beta_2 R1 + \beta_3 R1^2 + \beta_4(1 - z1)R1 + \beta_5(1 - z1)R1^2 \\
& + \beta_6(1 - z1)R1' + \beta_7(1 - z1)R1 \cdot R1' + \beta_8(1 - z1)R1^2 \cdot R1' \\
& + \beta_9(1 - z1)g'(R1) + \beta_{10}(1 - z1)g''(R1),
\end{aligned} \tag{C.19}$$

where $g''(R1) = \int_{-\infty}^{0} x^2 \cdot \phi\left(\frac{\mu_2 + \rho(R1 - \mu_1) - x}{\sqrt{1 - \rho^2}}\right) dx$.

We then fit (C.19) and compute the Wald estimator as above: $\widehat{\psi}_{R1}^{ATF}(Wald) = \frac{\widehat{\beta_1}}{\widehat{\gamma_1}}$.

Figure C.1

Relationship Between Probability of Treatment and Rating 1



Appendix D

Domain or Bandwidth Selection

This appendix first describes how we implemented bandwidth selection for each of three estimation methods (the frontier, fuzzy IV, and binding-score methods) and then presents bandwidth selection results.

D.1 Bandwidth Selection Algorithms

Here we provide a step-by-step description of our bandwidth selection algorithm for each of the three MRRDD estimation methods. For parsimony, we describe how these algorithms were used to select an optimal bandwidth around the frontier for rating 1 (ψ_{R1}^{ATF}).[1] Following the literature, we estimated local linear regression models on each side of the cut-point. For these regressions we weighted all sample members within the bandwidth equally, rather than choosing a more complex weighting function or kernel, which has not made much difference in practice (Imbens and Lemieux, 2008a). Our bandwidth selection goal was to identify a bandwidth that, given our choice of a linear functional form, provided the best possible estimator of the mean treatment effect at a frontier. Since this treatment effect equals the difference between the average outcomes at the frontiers with and without treatment, the optimal bandwidth is the one that yields the best predictions of these two conditional averages. Below we provide a step-by-step description of the bandwidth selection algorithm for each of the three MRRDD estimation methods that were considered.

Bandwidth selection for the frontier method. The frontier method uses only those sample points for which treatment assignment is fully determined by a single rating. For example, when focusing on the $R1$ frontier, we only used sample points where $R2 \geq c2$. Because we centered each rating on its cut-point value, this implies that we only used sample points where $R2 \geq 0$.

1. The first step entailed determining candidate bandwidths, which were defined in terms of their *width* (e.g., the interval of $R1$ that spans the bandwidth). For this purpose we specified candidate bandwidths that represent the following vector of percentiles ($P = (0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.75, 0.90)$). The algorithm then began by identifying the side of the frontier with the smaller number of sample points and established candidate bandwidths on that side.[2] This process

[1] By reversing the roles of the two ratings, the same process can be used to select an optimal bandwidth around the frontier for rating 2.

[2] Assume, for example, that fewer sample points were on the left side of the $R1$ frontier than were on the right side. In this case, the left side ($R1 \leq 0$ and $R2 \geq 0$) would be used to establish candidate bandwidths. Now assume that that the 5th-percentile value of $R1$ on the left side of the frontier was -2. The corresponding value on the right side of the cut-point would therefore be +2. Another strategy for defining candidate bandwidths (which we did not use) is to choose the 5th-percentile value of $R1$ on the left side of its cut-point and the 5th-percentile value of $R1$ on the right side of its cut-point.

was repeated to identify candidate bandwidths on the left and right sides of the cut-point for each percentile.

2. The next step entailed identifying a "range" of criterion sample points near the frontier, which were used to assess how well (in terms of mean squared error) a linear model can predict points at or near the frontier. We set the range to include the 100 sample points that were closest to the frontier on the left and the 100 sample points that were closest to the frontier on the right.

3. For all candidate bandwidths the algorithm proceeded as follows: On each side of the cut-point 100 local linear regressions of the form $Y_i = \alpha_0 + \alpha_1 R1_i + \epsilon_i$ were fit for each of the 100 sample points in the range.[3] Next, for each point in the range (for each of 200 regressions), a predicted value was computed given the estimated linear model and the value of the rating for that sample point.

4. For each candidate bandwidth, the algorithm then computed the resulting mean squared error (MSE) using the predicted and actual outcomes for the 200 sample points in the range. The candidate bandwidth with the lowest MSE was selected as the optimal bandwidth for a linear model. Note that this bandwidth was the same on both sides of the cut-point.[4]

Bandwidth selection for the fuzzy IV method. The fuzzy IV method uses sample points from all four quadrants to estimate average frontier-specific effects based on a first-stage model and a reduced-form model. The first-stage model specifies receiving treatment offer (*T*) as a function of rating 1 (*R*1) and an instrument (*z*1) that indicates assignment to treatment or the control group according to the value of $R1$. The reduced-form model specifies an outcome measure (Y) as a function of $R1$ and $z1$. It is therefore necessary to consider an optimal bandwidth for both models. The first two steps of bandwidth selection (choosing candidate bandwidths and the range) were the same as those for the frontier method. Then for each candidate bandwidth, the algorithm computed a separate MSE for the first-stage IV model and the reduced-form IV model as described below.

1. For the *first-stage IV model*, the value of T was always equal to one on the left side of the frontier because all sample members who scored below the rating 1 cut-point were assigned to treatment. Consequently, bandwidth selection was only necessary

[3] For example, if the candidate bandwidth was two units wide, then on the left side of the cut-point, linear regressions were fit using data for all sample points that were within two units to the left of each sample point in the range on the left, and on the right side of the cut-point, linear regressions were fit using data for all sample points that were within two units to the right of each sample point in the range on the right.

[4] Another option, which we did not explore, is to select a separate optimal bandwidth on each side of the frontier.

on the right side of the cut-point. On this side, a linear regression of the form $T_i = \gamma_0 + \gamma_1 R1_i + u_i$ was fit for each of the 100 sample points in the range. The regression for each sample point in the range was fit using all sample points that were within one candidate bandwidth to the right of it. The squared difference between the actual and predicted values of T for the sample points in the range was then used to compute the first-stage MSE for each candidate bandwidth.

2. For the *reduced-form IV model*, just like the frontier method, the algorithm used data on both sides of the $R1$ cut-point, only here it considered the entire sample, not just a partition of the sample. A local linear regression of the form $Y_i = \alpha_0 + \alpha_1 R1_i + \epsilon_i$ was fit for each of the 100 sample points in the range on the left side of the cut-point and each of the 100 sample points on the right side of the cut-point. The difference between the actual and predicted values of Y for these 200 points was used to compute the reduced-form MSE for each candidate bandwidth.

We compared the optimal bandwidths for the first-stage model and the reduced-form model and chose the smaller of the two for estimating program effects. This made it possible to estimate program effects using a common sample for both stages of IV estimation.[5]

Bandwidth selection for the binding-score method. Recall that the binding-score method devolves to a single-rating RDD, with the value of each sample member's single rating set equal to the minimum value of its multiple ratings. This made it possible to include all sample members in the analysis. Thus bandwidth selection for the binding-score method was the same as that for the frontier method except that the binding-score method used data for the full study sample. To do so we fit a model of the following form on both sites of the cut-point: $Y_i = \alpha_0 + \alpha_1 \min(R1_i, R2_i) + \epsilon_i$ as the basis for determining the MSE of candidate bandwidths.

D.2 Bandwidth Selection Results

Figures D.1 and D.2 depict bandwidth selection results. The three plots in Figure D.1 illustrate findings for the three correlations between $R1$ and $R2$, and the three plots in Figure D.2 illustrate findings for the three combinations of cut-points for $R1$ and $R2$. Each plot presents results for Models 1-4. Findings for Models 1 and 2 (homogeneous treatment effects) are reported for the frontier method, the fuzzy IV method, and the binding-score method. Findings for Models 3 and 4 (heterogeneous treatment effects) are not reported for the binding-

[5] It is possible to estimate the two stages of the IV model with different samples, but we did not explore this option.

score method. The figures report findings for estimating ψ_{R1}^{ATF}; corresponding findings for ψ_{R2}^{ATF} tell a similar story.

For each combination of model specification, estimation method, and correlation or cut-point of the ratings, the figures report a bar that depicts the percentage of selection of each candidate bandwidth across the 500 simulated samples. Bandwidths are ordered from the largest (top of each bar) to the smallest (bottom of each bar). The selection percentage for the smallest bandwidth (5th percentile) is shaded in white, and selection percentages for increasingly larger bandwidths are gradually shaded more darkly, with the largest bandwidth (90th percentile) shaded in black. Therefore, predominantly light bars point to the frequent selection of small bandwidths while predominantly dark bars indicate that large bandwidths were selected more frequently.

The first bar in each plot in Figures D.1 and D.2 is for the frontier method. When the data-generating models are linear in rating 1 (Models 1 and 3), these plots indicate that a large bandwidth is chosen more frequently. This is because a local linear regression matches the functional form of these data-generating models relatively well.[6] When the data-generating models are nonlinear (Models 2 and 4) small bandwidths are chosen more frequently than linear models. This is because the algorithm begins to omit the sample points that are farthest from the frontier as it searches for the bandwidth for which a linear approximation is as good as possible, without sacrificing too much precision.

The second bar in each plot in Figures D.1 and D.2 is for the fuzzy IV method. These bandwidths are defined as the smaller of the two bandwidths for the first-stage and reduced-form IV models. We see that for all models, smaller bandwidths were chosen more frequently for the fuzzy IV method than for the other two methods. This might occur because the first-stage model of the IV method requires smaller bandwidths than the other methods or because the fuzzy IV method chooses the smaller of two bandwidths and thus has two shots at getting a small bandwidth. Nonetheless, optimal bandwidths varied substantially for the fuzzy IV method.

The third bar in each plot in Figures D.1 and D.2 (for Models 1 and 2 only) presents bandwidth selection results for the binding-score method. These plots indicate that a linear approximation of the relationship between sample members' outcomes and binding scores is a good one for the linear data-generating Model 1. Thus large bandwidths were chosen fre-

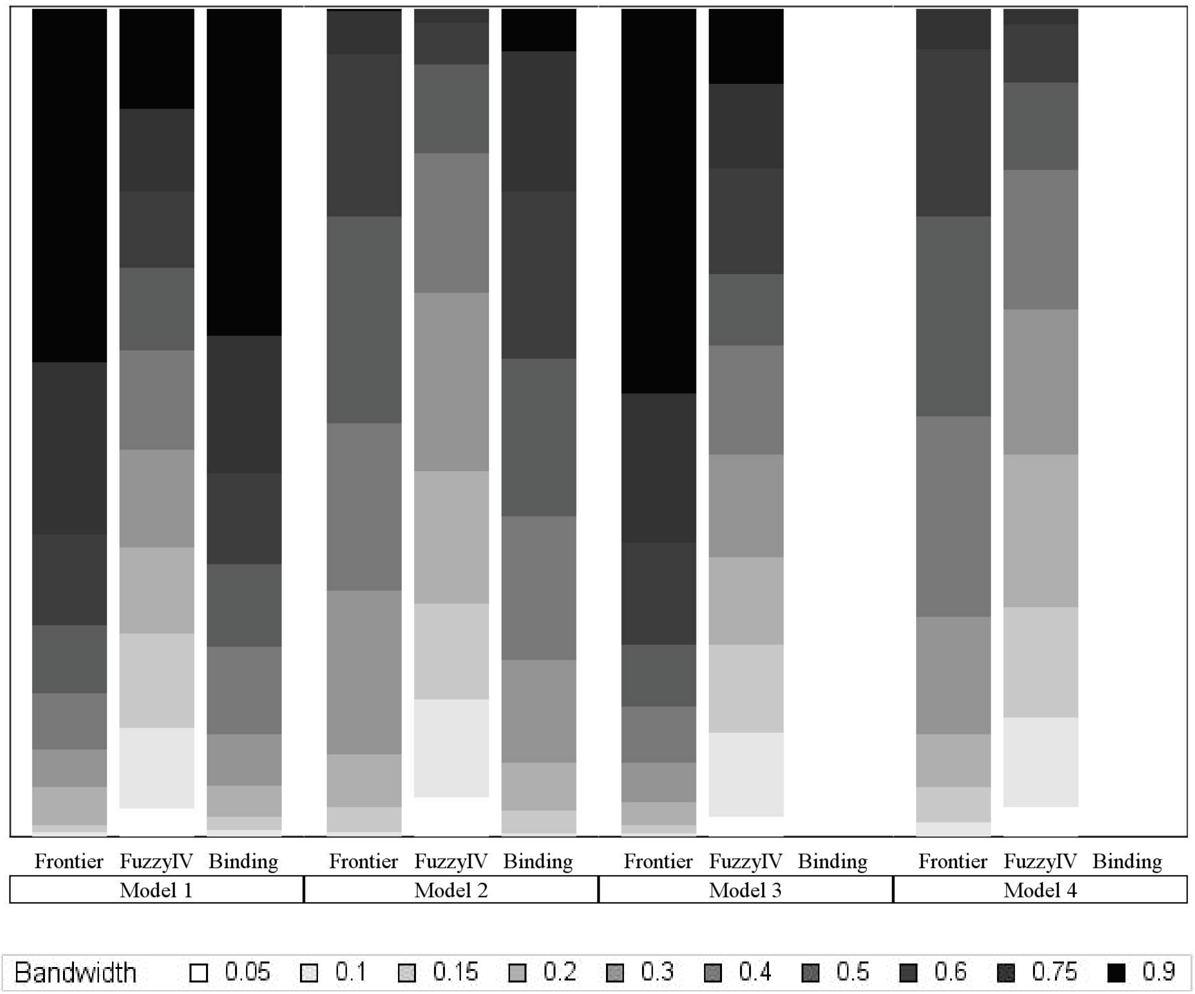
[6] Note that the fact that the underlying data-generating model is linear in $R1$ does not imply that $E(Y|R1, R2 \geq 0)$ is linear, because the latter depends on the joint distribution of $R1$ and $R2$ in the region where $R2 \geq 0$. It is possible to construct other data-generating models where Y is linear in $R1$ but where the function $f(R1) = E(Y|R1, R2 \geq 0)$ is highly nonlinear in $R1$.

quently for this model. Large bandwidths were selected less frequently for Model 2, which is quadratic in $R1$.

Overall we see that a wide range of bandwidths are selected for all of the data-generating models and for each of estimation methods. However, larger bandwidths tend to be selected more frequently when the data are generated by a linear model, and the fuzzy IV method tends to use bandwidths that are smaller than those for the other two methods. Finally, very small bandwidths are rarely selected. This suggests that moderate to large bandwidths can reasonably approximate even nonlinear data-generating models.

Figure D.1

Bandwidth Selection Results When Using Local Linear Regression, by Correlation Between Ratings, All Data-Generating Models

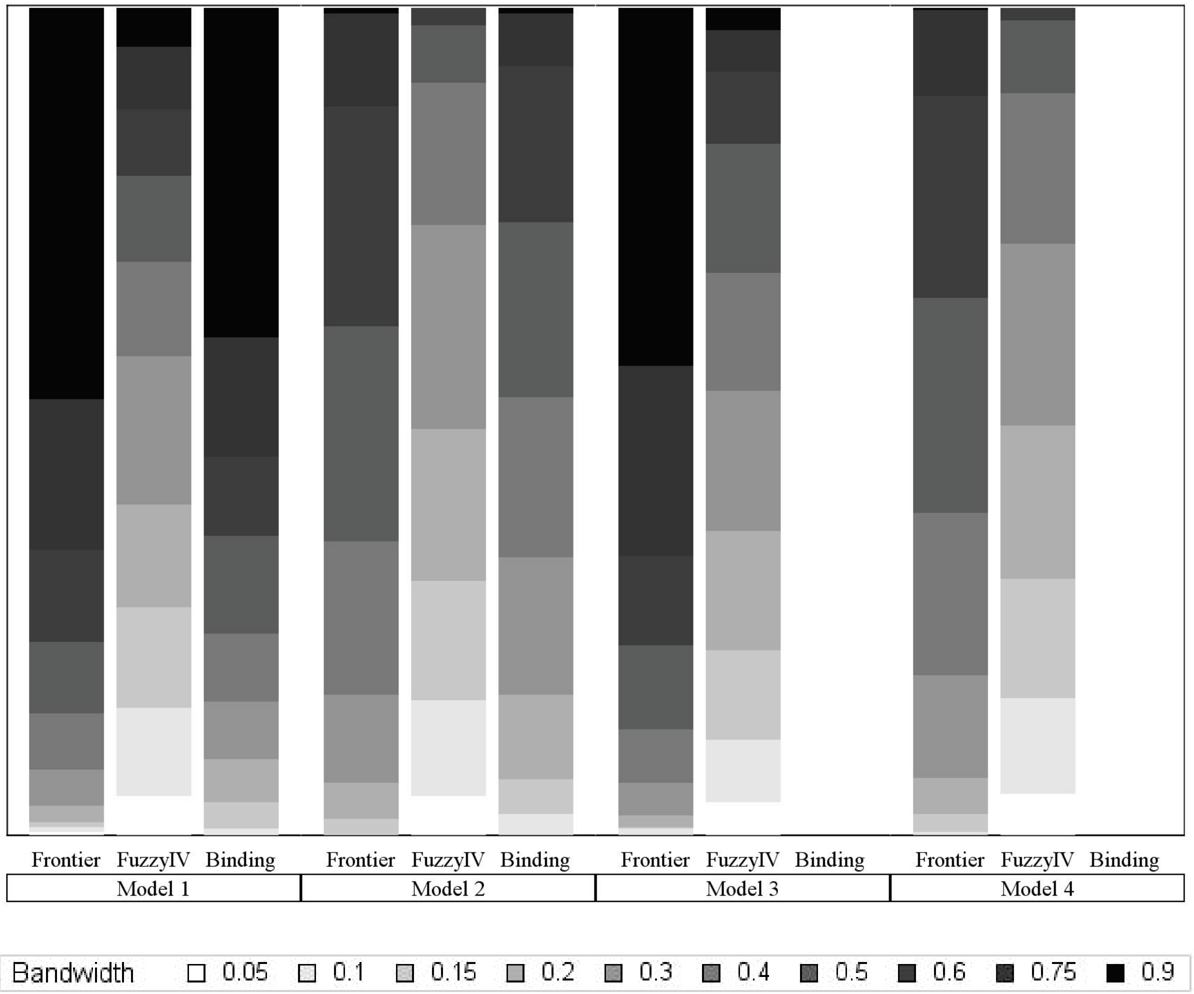


cor(R1,R2)=0.2, cuts=50/50

(continued)

Figure D.1 (continued)

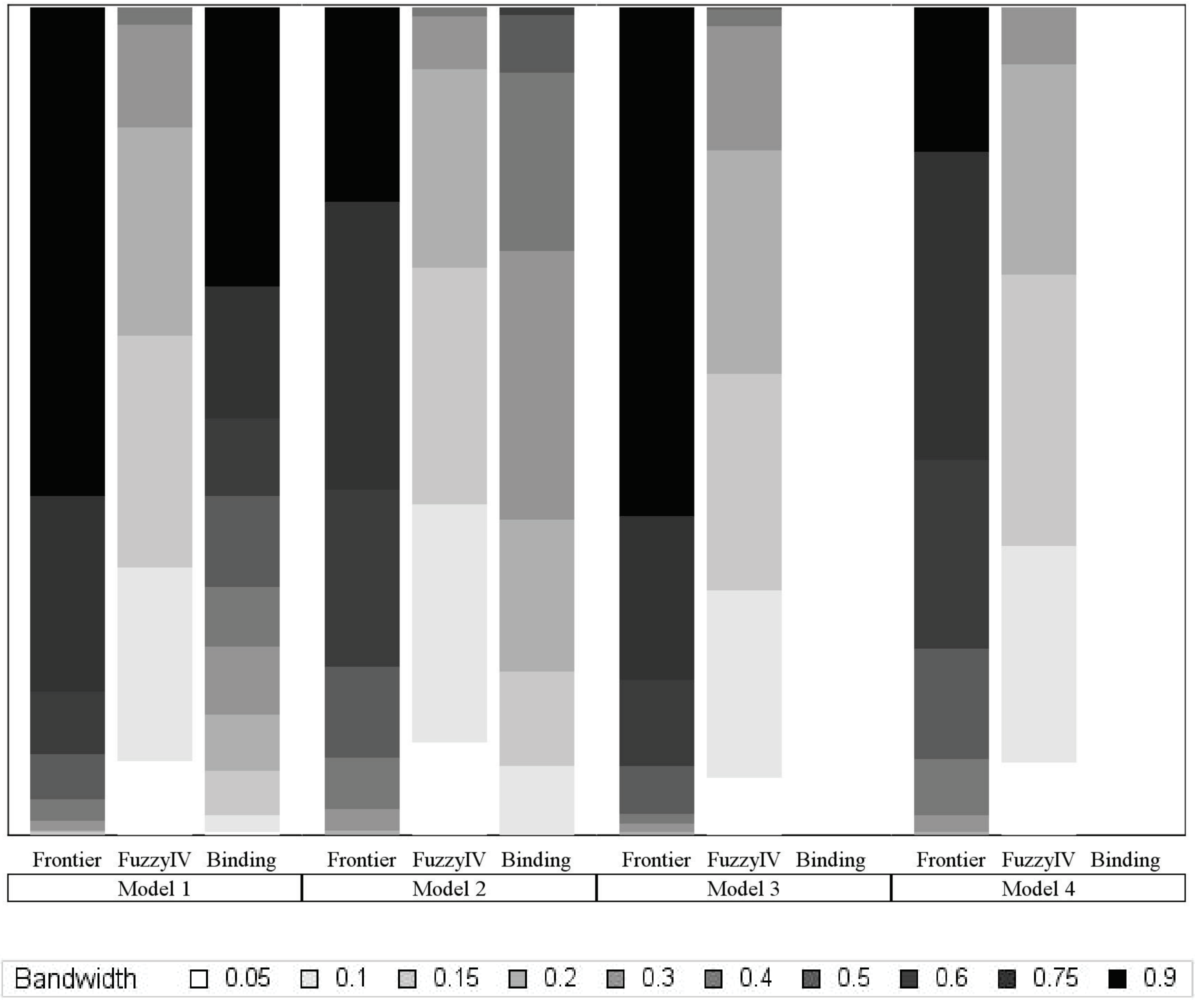
cor(R1,R2)=0.5, cuts=50/50



(continued)

Figure D.1 (continued)

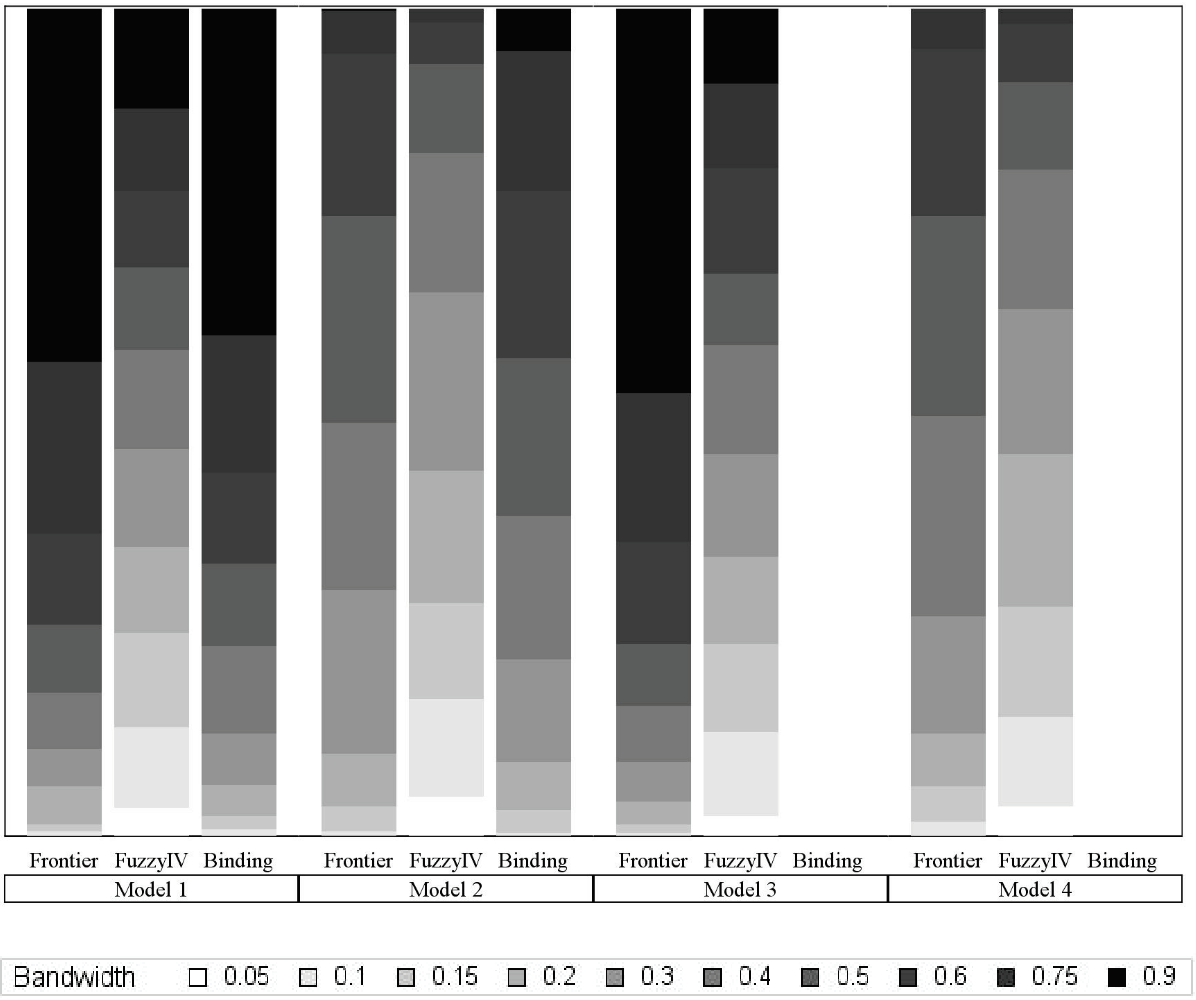
cor(R1,R2)=0.9, cuts=50/50



NOTE: For 500 samples of size 5,000.

Figure D.2

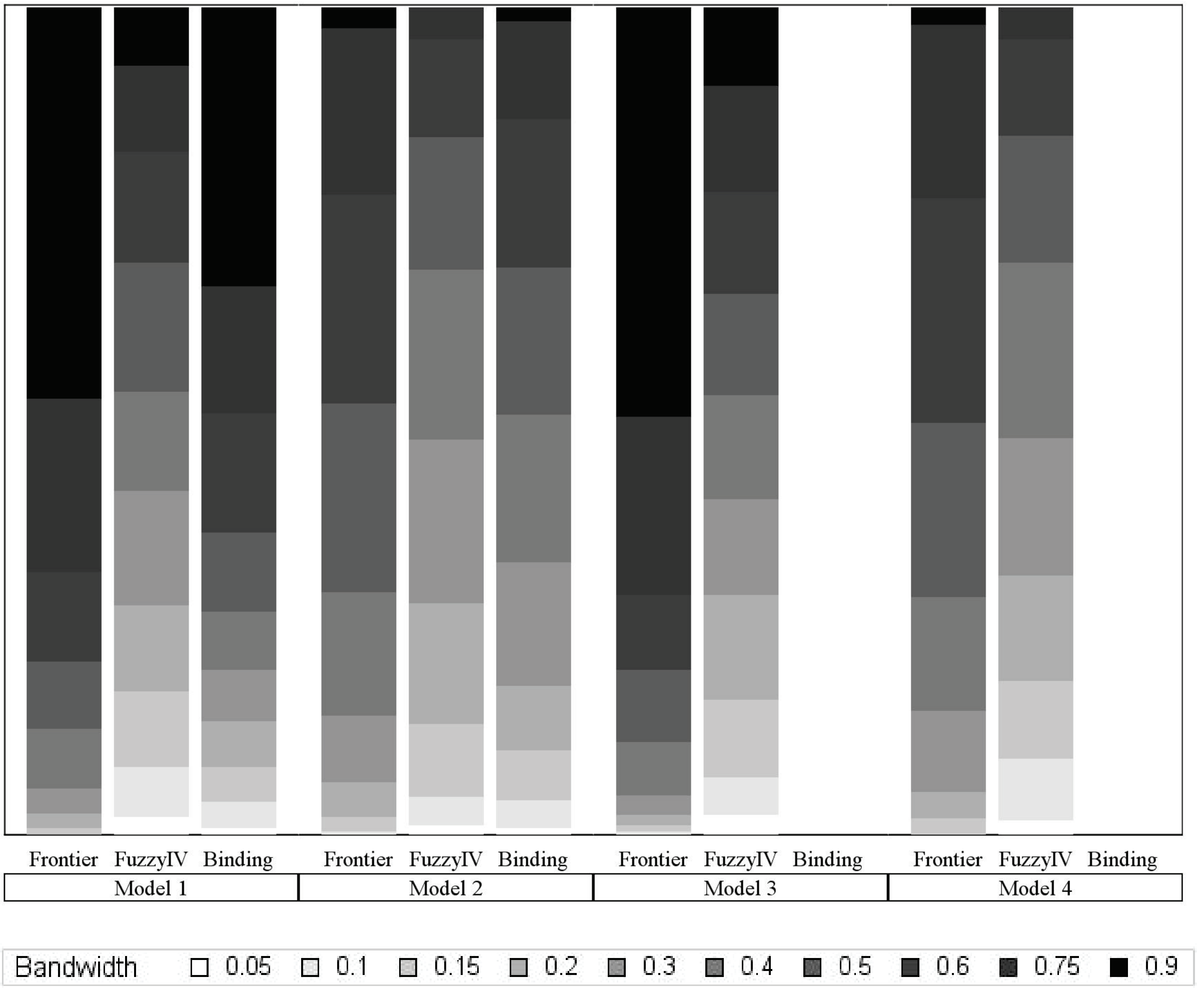
Bandwidth Selection Results When Using Local Linear Regression, by Location of Cut-Points in Ratings, All Data-Generating Models



cor(R1,R2)=0.2, cuts=50/50

(continued)

Figure D.2 (continued)

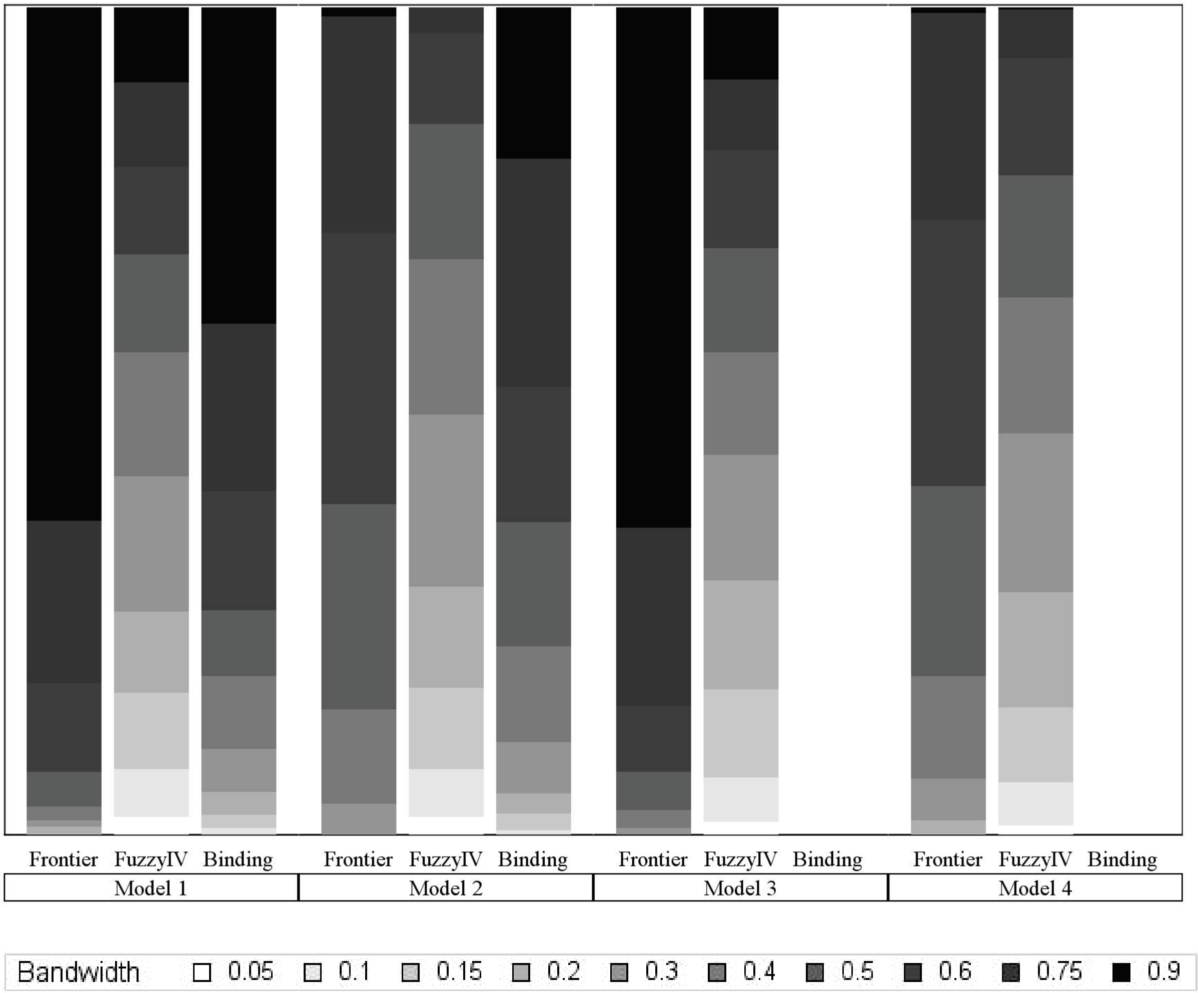


cor(R1,R2)=0.2, cuts=30/30

(continued)

Figure D.2 (continued)

cor(R1,R2)=0.2, cuts=30/70



NOTE: For 500 samples of size 5,000.

Appendix E

Simulation Results When Using Local Linear Regression, Not Including Other Rating as a Covariate

Table E.1

Simulation Results When Using Local Linear Regression, Not Including Other Rating as a Covariate, Data-Generating Model 1

						Mean			
						Estimate	Relative		Relative
Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	of $Var(\hat{\psi})$	Var[a]	MSE	MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.403	0.003	0.025	0.015	1.000	0.025	1.000
		Fuzzy IV	0.420	0.020	0.157	0.106	6.253	0.157	6.266
		Binding	0.399	-0.001	0.022	0.011	0.865	0.022	0.864
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.401	0.001	0.016	0.012	1.000	0.016	1.000
		Fuzzy IV	0.391	-0.009	0.095	0.061	5.893	0.095	5.898
		Binding	0.399	-0.001	0.022	0.011	1.348	0.022	1.348
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.399	-0.001	0.031	0.017	1.000	0.031	1.000
		Fuzzy IV	0.424	0.024	0.145	0.111	4.660	0.146	4.678
		Binding	0.396	-0.004	0.017	0.010	0.542	0.017	0.542
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.389	-0.011	0.021	0.013	1.000	0.021	1.000
		Fuzzy IV	0.400	0.000	0.099	0.073	4.740	0.099	4.711
		Binding	0.396	-0.004	0.017	0.010	0.807	0.017	0.803
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.401	0.001	0.032	0.023	1.000	0.032	1.000
		Fuzzy IV	0.423	0.023	0.141	0.126	4.422	0.141	4.438
		Binding	0.400	0.000	0.017	0.009	0.536	0.017	0.536
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.399	-0.001	0.032	0.021	1.000	0.032	1.000
		Fuzzy IV	0.401	0.001	0.129	0.109	4.024	0.129	4.024
		Binding	0.400	0.000	0.017	0.009	0.532	0.017	0.532
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.394	-0.006	0.018	0.015	1.000	0.018	1.000
		Fuzzy IV	0.417	0.017	0.127	0.082	6.906	0.128	6.905
		Binding	0.399	-0.001	0.016	0.010	0.865	0.016	0.863
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.409	0.009	0.021	0.013	1.000	0.022	1.000
		Fuzzy IV	0.410	0.010	0.067	0.055	3.114	0.067	3.108
		Binding	0.399	-0.001	0.016	0.010	0.744	0.016	0.741
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.399	-0.001	0.038	0.028	1.000	0.038	1.000
		Fuzzy IV	0.561	0.161	6.533	1.252	172.072	6.559	172.750
		Binding	0.394	-0.006	0.018	0.012	0.465	0.018	0.466
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.398	-0.002	0.019	0.012	1.000	0.019	1.000
		Fuzzy IV	0.391	-0.009	0.048	0.040	2.519	0.048	2.523
		Binding	0.394	-0.006	0.018	0.012	0.933	0.018	0.935

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a]Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table E.2

Simulation Results When Using Local Linear Regression, Not Including Other Rating as a Covariate, Data-Generating Model 2

Cuts/ρ	Parameter	Method	E($\hat{\psi}$)	Bias	Var($\hat{\psi}$)	Mean Estimate of Var($\hat{\psi}$)	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.383	-0.017	0.116	0.085	1.000	0.117	1.000
		Fuzzy IV	0.294	-0.106 **	0.288	0.241	2.476	0.300	2.567
		Binding	0.375	-0.025	0.132	0.086	1.135	0.133	1.137
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.399	-0.001	0.219	0.141	1.000	0.219	1.000
		Fuzzy IV	0.276	-0.124 **	0.770	0.455	3.517	0.786	3.588
		Binding	0.375	-0.025	0.132	0.086	0.603	0.133	0.606
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.415	0.015	0.084	0.069	1.000	0.084	1.000
		Fuzzy IV	0.330	-0.070 **	0.224	0.194	2.674	0.229	2.726
		Binding	0.370	-0.030 *	0.095	0.065	1.135	0.096	1.143
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.375	-0.025	0.167	0.109	1.000	0.167	1.000
		Fuzzy IV	0.290	-0.110 **	0.440	0.376	2.638	0.452	2.700
		Binding	0.370	-0.030 *	0.095	0.065	0.571	0.096	0.574
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.417	0.017	0.042	0.035	1.000	0.043	1.000
		Fuzzy IV	0.376	-0.024	0.136	0.137	3.230	0.137	3.220
		Binding	0.401	0.001	0.028	0.021	0.673	0.028	0.668
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.431	0.031 **	0.046	0.036	1.000	0.047	1.000
		Fuzzy IV	0.387	-0.013	0.230	0.151	4.980	0.230	4.880
		Binding	0.401	0.001	0.028	0.021	0.617	0.028	0.604
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.408	0.008	0.127	0.110	1.000	0.127	1.000
		Fuzzy IV	0.335	-0.065	0.272	0.236	2.146	0.276	2.178
		Binding	0.393	-0.007	0.175	0.108	1.380	0.175	1.379
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.365	-0.035	0.343	0.207	1.000	0.344	1.000
		Fuzzy IV	0.338	-0.062	0.819	0.535	2.391	0.823	2.393
		Binding	0.393	-0.007	0.175	0.108	0.509	0.175	0.508
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.400$	Frontier	0.403	0.003	0.112	0.105	1.000	0.112	1.000
		Fuzzy IV	0.256	-0.144 **	1.437	1.589	12.844	1.457	13.027
		Binding	0.399	-0.001	0.273	0.179	2.438	0.273	2.437
	$\psi^{ATF}_{R2} = 0.400$	Frontier	0.405	0.005	0.333	0.216	1.000	0.333	1.000
		Fuzzy IV	0.358	-0.042	0.660	0.520	1.982	0.662	1.987
		Binding	0.399	-0.001	0.273	0.179	0.819	0.273	0.819

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a]Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table E.3

Simulation Results When Using Local Linear Regression, Not Including Other Rating as a Covariate, Data-Generating Model 3

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Mean Estimate of $Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.244$	Frontier	0.248	0.004	0.023	0.014	1.000	0.023	1.000
		Fuzzy IV	0.260	0.016	0.109	0.085	4.826	0.110	4.835
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.322$	Frontier	0.331	0.009	0.018	0.012	1.000	0.018	1.000
		Fuzzy IV	0.327	0.006	0.079	0.062	4.525	0.079	4.506
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.262$	Frontier	0.272	0.010	0.026	0.015	1.000	0.026	1.000
		Fuzzy IV	0.270	0.008	0.118	0.085	4.504	0.118	4.489
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.331$	Frontier	0.332	0.002	0.018	0.012	1.000	0.018	1.000
		Fuzzy IV	0.355	0.025	0.097	0.065	5.364	0.097	5.397
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.330$	Frontier	0.330	-0.001	0.028	0.022	1.000	0.028	1.000
		Fuzzy IV	0.336	0.006	0.113	0.111	4.042	0.113	4.044
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.365$	Frontier	0.376	0.011	0.027	0.021	1.000	0.027	1.000
		Fuzzy IV	0.359	-0.006	0.112	0.112	4.157	0.112	4.141
		Binding	NA	NA	NA	NA	NA	NA	NA
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.209$	Frontier	0.203	-0.007	0.026	0.014	1.000	0.026	1.000
		Fuzzy IV	0.205	-0.005	0.113	0.071	4.430	0.113	4.423
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.304$	Frontier	0.304	0.000	0.020	0.012	1.000	0.020	1.000
		Fuzzy IV	0.293	-0.011	0.057	0.044	2.780	0.057	2.786
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.281$	Frontier	0.279	-0.002	0.034	0.026	1.000	0.034	1.000
		Fuzzy IV	0.233	-0.048	8.553	2.312	249.174	8.556	249.199
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.294$	Frontier	0.292	-0.002	0.020	0.011	1.000	0.020	1.000
		Fuzzy IV	0.294	0.000	0.057	0.040	2.794	0.057	2.794
		Binding	NA	NA	NA	NA	NA	NA	NA

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a] Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

Table E.4

Simulation Results When Using Local Linear Regression, Not Including Other Rating as a Covariate, Data-Generating Model 4

Cuts/ρ	Parameter	Method	$E(\hat{\psi})$	Bias	$Var(\hat{\psi})$	Mean Estimate of $Var(\hat{\psi})$	Relative Var[a]	MSE	Relative MSE[a]
Correlation between ratings varying, cut-points held constant									
ρ = 0.20 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.167$	Frontier	0.152	-0.014	0.116	0.083	1.000	0.117	1.000
		Fuzzy IV	0.109	-0.057 *	0.257	0.219	2.203	0.260	2.227
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.245$	Frontier	0.238	-0.008	0.181	0.127	1.000	0.181	1.000
		Fuzzy IV	0.182	-0.063	0.586	0.486	3.241	0.590	3.261
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.50 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.201$	Frontier	0.209	0.007	0.086	0.067	1.000	0.086	1.000
		Fuzzy IV	0.119	-0.083 **	0.218	0.179	2.523	0.225	2.603
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.271$	Frontier	0.259	-0.011	0.142	0.103	1.000	0.142	1.000
		Fuzzy IV	0.219	-0.052	0.497	0.378	3.500	0.500	3.515
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.90 cut1 = 0.50 cut2 = 0.50	$\psi^{ATF}_{R1} = 0.315$	Frontier	0.340	0.025 **	0.047	0.035	1.000	0.048	1.000
		Fuzzy IV	0.318	0.003	0.128	0.127	2.719	0.128	2.683
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.350$	Frontier	0.356	0.006	0.046	0.039	1.000	0.046	1.000
		Fuzzy IV	0.311	-0.039 *	0.158	0.148	3.402	0.159	3.431
		Binding	NA	NA	NA	NA	NA	NA	NA
Cut-points varying, correlation between ratings held constant									
ρ = 0.20 cut1 = 0.30 cut2 = 0.30	$\psi^{ATF}_{R1} = 0.100$	Frontier	0.104	0.004	0.143	0.099	1.000	0.144	1.000
		Fuzzy IV	0.069	-0.031	0.341	0.247	2.376	0.342	2.383
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.196$	Frontier	0.238	0.042	0.261	0.184	1.000	0.262	1.000
		Fuzzy IV	0.128	-0.068 *	0.532	0.452	2.041	0.537	2.045
		Binding	NA	NA	NA	NA	NA	NA	NA
ρ = 0.20 cut1 = 0.30 cut2 = 0.70	$\psi^{ATF}_{R1} = 0.235$	Frontier	0.261	0.025	0.154	0.107	1.000	0.155	1.000
		Fuzzy IV	0.259	0.024	1.406	1.243	9.130	1.406	9.096
		Binding	NA	NA	NA	NA	NA	NA	NA
	$\psi^{ATF}_{R2} = 0.164$	Frontier	0.160	-0.004	0.306	0.204	1.000	0.306	1.000
		Fuzzy IV	0.103	-0.061	0.675	0.535	2.205	0.679	2.217
		Binding	NA	NA	NA	NA	NA	NA	NA

NOTES: For 500 samples of size 5,000.

Statistical significance levels are indicated as follows: ** = 1 percent; * = 5 percent.

[a] Relative variance and relative MSE for given estimation method divided by the variance or MSE of the frontier estimation method.

References

Black, D. A., Galdo, J., and Smith, J. A. (2007). Evaluating the Worker Profiling and Reemployment Services System Using a Regression Discontinuity Approach. *The American Economic Review, 97*(2), 104-107.

Bloom, H. S. (2012). Modern Regression Discontinuity Analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43-82.

Cook, T. D., Shadish, W. R., and 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*, 724-750.

Gamse, B. C., Bloom, H. S., Kemple, J. J., and Jacob, R. T. (2008). Reading First Impact Study: Interim Report. NCEE 2008-4016: National Center for Education Evaluation and Regional Assistance. Available from: ED Pubs. P.O. Box 1398, Jessup, MD 20794-1398. Tel.: 877-433-7827; website: http://ies.ed.gov/ncee/pubs/.

Gill, B., Lockwood, J. R., Martorell, F., Setodji, C. M., and Booker, K. (2008). State and Local Implementation of the No Child Left Behind Act. In O. o. P. U.S. Department of Education, Evaluation and Policy Development, Policy and Program Studies Service (Ed.).

Goldberger, A. S. (1972a). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*. University of Wisconsin, Institute for Research on Poverty, June. Discussion Paper 126-72, Madison, WI.

Goldberger, A. S. (1972b). *Selection Bias in Evaluating Treatment Effects: The Case of Interaction*. University of Wisconsin, Institute for Research on Poverty. Madison, WI.

Hahn, J., Todd, P., and Van Der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression Discontinuity Design. *Econometrica, 69*, 201-209.

Imbens, G., and Kalyanaraman, K. (2009). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. NBER Working Paper No. 14726. Retrieved from http://www.nber.org/papers/w14726.

Imbens, G., and Lemieux, T. (2008a). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics, 142*(2), 615-635.

Imbens, G., and Lemieux, T. (2008b). Special Issue: The Regression Discontinuity Design — Theory and Applications. *Journal of Econometrics, 142*(2).

Jacob, B. A., and Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics, 86*(1), 226-244.

Kane, T. J. (2003). *A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going*. NBER Working Paper No. 9703. Retrieved from http://www.nber.org/papers/w9703.

Lee, D. S. (2008). Randomized Experiments from Non-Random Selection in U.S. House Elections. *Journal of Econometrics, 142*, 675-697.

Lee, D. S., and Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, XLVIII*, 281-355.

Ludwig, J., and Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Approach. *Quarterly Journal of Economics 122*(1), 159-208. doi: 10.1162/qjec.122.1.159

Martorell, P. (2005). Do High School Graduation Exams Matter? Evaluating the Effects of Exit Exam Performance on Student Outcomes. Unpublished working paper, University of California, Berkeley.

Matsudaira, J. D. (2008). Mandatory Summer School and Student Achievement. *Journal of Econometrics, 142*.

Mosteller, F. (1990). Improving Research Methodology: An Overview. In L. Sechrest, E. Perrin, and J. Bunker (eds.), *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data* (pp. 221-230). Rockville, MD: U.S. Public Health Service, Agency for Health Care Policy and Research.

Niu, S. X., and 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), 89-110.

Ou, D. (2009). To Leave or Not to Leave? A Regression Discontinuity Analysis of the Impact of Failing High School Exit Exam. CEE Discussion Papers 0107, Centre for the Economics of Education, London School of Economics.

Papay, J. P., Murnane, R. J., and 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.

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011). Extending the Regression Discontinuity Approach to Multiple Assignment Variables: Evidence from the Massachusetts High School Exit Examination. *Journal of Econometrics, 161*(2), 203-207.

Reardon, S. F., Arshan N., Atteberry, A. and Kurlaender, M. (2010). Effects of Failing a High School Exit Exam on Course-Taking, Achievement, Persistence, and Graduation. *Education Evaluation and Policy Analysis, 32*(4), 498-520.

Reardon, S. F., and Robinson, J. P. (2012). Regression Discontinuity Designs with Multiple Rating-Score Variables. *Journal of Research on Educational Effectiveness, 5*(1), 83-104.

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. (2011). Evaluating Criteria for English Learner Reclassification: A Causal-Effects Approach Using a Binding-Score Regression Discontinuity Design with Instrumental Variables. *Educational Evaluation and Policy Analysis, 33*, 267–292.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics, 2*(1), 1-26.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., and Smith, J. (2010). Standards for Regression Discontinuity Designs. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Weinbaum, E. H., and Weiss, M. J. (2009). School Response to NCLB Labels. Paper presented at the annual meeting of the American Education Research Association, San Diego, CA.

Wong, V. C., Steiner, P. M., and Cook, T. D. (2013). Analyzing Regression-Discontinuity Designs with Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. *Journal of Educational and Behavioral Statistics, 38*(2), 107–141. doi: 10.3102/1076998611432172

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.