Evaluating the Regression Discontinuity Design Using Experimental Data

Dan Black
University of Chicago and NORC
danblack@uchicago.edu

Jose Galdo
McMaster University and IZA
galdojo@mcmaster.ca

Jeffrey Smith
University of Michigan, NBER, IFS, IZA, PSI and ZEW
econjeff@umich.edu

Keywords: Regression-Discontinuity, Cross-Validation, Kernel Regression**.**
JEL Classification: C13, C14.

Version: April 2007

We thank along with Josh Angrist, Chris Taber, and Wilbert Van der Klaauw for their comments and useful discussions. We also thank seminar participants at Penn University, Queen's University, and Laval University, and participants in the 2005 Econometric Society World Congress, the CREST-INSEE Conference on "Econometric Evaluation of Public Policies", and the 2006 Society of Labor Economists Meeting for helpful comments and suggestions.

Evaluating the Regression Discontinuity Design Using Experimental Data

Abstract

The regression discontinuity (RD) design has recently become a standard method for identifying causal effects for policy interventions. We use an unusual "tie breaking" experiment, the Kentucky Working Profiling and Reemployment Services, to investigate the performance of widely used RD estimators. Two features characterize this program. First, the treatment (reemployment services) is assigned as a discontinuous function of a profiling variable (expected benefit receipt duration), which allows the identification of both experimental and nonexperimental samples. Second, we deal with a discontinuity frontier rather than a discontinuity point, which allows the identification of local average treatment effects over a wide range of the support of the discontinuous variable. Using a variety of multivariate parametric and nonparametric kernel estimators, we estimate the bias with respect to the benchmark experimental estimates. In general, we find that local linear kernel estimates show the least bias, but parametric estimates perform reasonably well. We also examine two alternative discontinuities – geography and time – and find that they provide credible estimates as well.

I. Introduction

The regression-discontinuity (hereafter RD) design has recently become a standard evaluation framework for solving causal issues with nonexperimental data. The intrinsic feature of this approach is there is jump in an increase in the probability of treatment when an observed covariate crosses a known threshold (Trochim 1984). This design allows one to identify the program's causal effect without imposing exclusion restrictions, index assumptions on the selection process, functional forms, or distributional assumptions on errors; see Hahn, Todd, and Van der Klaauw (2001).

The empirical literature applying the RD approach can be traced back to Thistlethwaite and Campbell's (1960) pioneering work that estimated the effect that receipt of a National Merit Award has on a student's later success. As the award is given to students who achieve a minimum score, differences in future academic achievement between those students above and below that cut-off is attributed to the effect of the award. More recently, the evaluation literature has shown a renewed interest in exploiting information about discontinuities in the treatment assignment. Hahn et al. (2001) is the first to link RD design to the program evaluation literature and to formally establish weaker conditions for identification. Lee (2005) argues that RD mimics random assignment of treatment status near the cut-off), and Porter (2004) studies issues involved in nonparametric estimation of treatment impacts at the discontinuity points. Recent empirical applications include Angrist and Lavy (1999), Van der Klaauw (2002), DiNardo and Lee (2004), Lemieux and Milligan (2004), Chen and Van der Klaauw (2004), Martorell (2004), Matsudaira (2004), Black, Galdo, and Smith (2007), Card...(2007), Lalive (2006, 2007), and {paper from the ASSA session)xxxxxx

These empirical studies rely on observational data that prevents the evaluation of the performance of RD econometric estimators in solving the evaluation problem. In this study, we study the performance widely-used RD estimators following LaLonde (1986) and Fraker and Maynard (1986) who evaluates the performance of nonexperimental estimators using experimental data as benchmark. We exploit a unique tie-breaking experiment, the Kentucky Working and Profiling Reemployment Services (hereafter WPRS), that seeks behavioral effects on Unemployment Insurance (UI) claimants with expected high probabilities of benefit exhaustion. The Kentucky WPRS program employs a statistical model to estimate the expected duration of each new UI claim as a function of the claimants' characteristics and his or her local economic characteristics. In each local UI office and each week, new claimants are assigned to receive mandatory reemployment and training services based on their scores. Assignment starts with the high scores and continues until the number of slots available for a given office in a given week is reached. Within the marginal profiling score – the one at which capacity is reached – random assignment allocates the claimants into treatment. In Figure One, we depict four groups that might be used to evaluate the program each week. The random assignment forms the experimentally determined treated group (B) and control group (C). Black, Smith, Berger, and Noel (2003) provide experimental evidence of a large impact of the Kentucky WPRS program.

Two additional groups of individuals can be identified from observed discontinuities above and below the marginal scores. They are the nonexperimental treated group individuals (D) and comparison group individuals (A). Thus, this unique design allows us to exploit a series of "sharp" discontinuities in the assignment of the requirement to receive re-employment services inherent in the implementation of the WPRS program. It also allows for the identification of both experimental and nonexperimental data without the need for resorting to

"external" comparable groups. In this sense, one of the contributions of this paper is its reliance on high-quality data that place all experimental and nonexperimental treated and untreated individuals in the same local labor markets and under the same administrative surveys and questionnaires.

We assume the experimental estimates to be unbiased estimates of the true program effects. As a result, we interpret significant differences between these two sets of estimates are taken as evidence of the failure of the RD design to provide reliable econometric estimators of the program effects given the available data. We also examine two alternative discontinuities by looking at neighbors in two other dimensions: time and geography. Which of these counterfactuals better replicates the experimentally determined treatment effects is a question of methodological and substantive interest.

One important difference with respect to traditional RD designs that are based on a single discontinuity point is that the Kentucky WPRS program embodies multiple discontinuities, as matched treated and untreated individuals are located on both sides of the profiling score boundary for a given week and a given local office. (Card, YYYY, (2007) is also an exception.) This characteristic allows us to identify treatment effects over a wider range of the support of the discontinuity variable and to estimate heterogeneous treatment impacts that vary with the score over the range of values that have treated and untreated individuals.

We have three main findings. First, the local linear kernel estimator with cross-validated optimal bandwidths best replicates the experimental estimates among all RD estimators. This is a consistent result for all samples used in the estimation and all outcomes of interest. Second, the parametric models replicate the experimental estimates reasonable well. By restricting the nonexperimental samples to units increasingly closer to the discontinuity points, the parametric

approach mimics estimates from the kernel approach. It seems that operating with a caliper aids our parametric models to better approximate the unknown conditional mean function. Furthermore, the inclusion of a set of baseline covariates has marginal effects on the treatment estimates and it does not alter any of the previous findings. Third, we find evidence that geography and time discontinuities also provide credible estimates. This is an important result because it greatly increases the potential application of the RD design and demonstrates that RD design is a promising evaluation method even when using discontinuities that arise across geography and time.

The paper proceeds as follows. Section II presents the program and the data. In Section III, we discuss the identification strategy and parameters of interest. Section IV describes the empirical framework used in the estimation of the parameters of interest. Section V presents the RD bias and explores the sensitivity of the estimates to several robustness specifications. In Section VI, we describe alternative counterfactuals by looking at two alternative discontinuities (time and geography). The final section concludes.

II. The Program and the Data

II.1 Institutional Background

The distortions provide by the UI system are widely acknowledge among economists. The incentives motivate claimants to extend their unemployment spells beyond what they would be in the absence of UI benefits, either by subsidizing additional job search or by subsidizing the consumption of leisure. In November 1993, President Clinton signed into law the Unemployment Compensation Amendments of 1993, which requires states to launch Worker Profiling and Reemployment Systems (WPRS) in order to reduce the duration of unemployment spells for those with higher probabilities of exhausting the 26 weeks of UI benefits.

In June 1994, the Commonwealth of Kentucky was selected as a demonstration state for implementing the WPRS program, which identifies potential exhaustees of the UI benefits among new initial claimants, and then assigns them mandatory reemployment services such as job-training and job-search workshops early in their spell so they may continue receiving benefits. The services themselves can be view either as a valuable opportunity to learn new employment-related skills, or as an in-kind tax on the leisure of the UI claimants (Black, et al 2003).

The Center for Business and Economic Research (CBER) at the University of Kentucky took responsibility for designing a statistical model to estimate the expected duration of each new UI claim as a function of the claimant's personal characteristics and his or her local economic characteristics. The model was estimated by employing five years of claimant data obtained from the Kentucky unemployment insurance mainframe computer databases, supplemented with data from other administrative data sources.

Two main features distinguish the Kentucky model from most other profiling models implemented in other states. First, the dependent variable is not represented by a dichotomous variable of whether the claimant exhausted UI benefits, but rather used the fraction of benefits received as a continuous variable. Unlike the binary exhaustion variable, the fraction of benefits exhausted variable distinguishes claimants who use 1 week of UI benefits from claimants who use 25 weeks of UI benefits. Second, the Kentucky's statistical model relies on over 140 covariates, including personal characteristics, industry and occupation controls, and local economic and labor market conditions.[1] By contrast, the model for the State of Pennsylvania uses only eight covariates, while the model for Washington State, which is one of the largest State models, includes 26 covariates. With these rich data, the model yields significant gains in

[1] It is against the law to profiled based on ethnicity, age, sex, and veteran status.

predictive power with respect to profiling models from other states; see Black, Smith, Plesca, and Shannon (2003b) for details.

II.2 The KWPRS Treatment Assignment

The Kentucky model produces a single continuous measure of the fraction of UI benefits each claimant will collect. This profiling score is collapsed into a discrete score ranging from 1 to 20. Claimants predicted by the model to exhaust between 95 and 100 percent of their unemployment benefits receive a score of 20, claimants predicted to exhaust between 90 and 95 percent of their unemployment benefits receive 19, and so on. Figure 1 illustrates the Kentucky treatment assignment. For each local employment office in each week, claimants starting new spells are ranked by their profiling scores. Those individuals with the highest scores are the first to be assigned to receive mandatory employment and training services, and this process continues until the number of slots available for each office in each week is reached.[2] Those claimants selected to receive reemployment services are contacted via mail to inform them about their rights and responsibilities under the program. Importantly, the treatment consists of the requirement to receive reemployment services, not actual receipt of these services. As many selected claimants may leave the UI system before receiving services but after being required to receive services, the KWPRS treatment can be considered as the intent-to-receive-treatment.

If the maximum number of claimants to receive reemployment in a given local office and in a given week is reached, and there are two or more claimants receiving the same discrete profiling score, a random number generator assigns the appropriate number of claimants to treatment. Therefore, only claimants with *marginal* profiling scores - the one at which the capacity constraint is reached in a given week and in a given local office - are randomly assigned

[2] The number of slots is determined *ex-ante*. It is more or less constant across time subject to resignations, vacations, sick leaves, et cetera.

into experimental treated and control groups. Black et al. (2003a) call these marginal sets of claimants "profiling tie groups", or PTGs. Finally, those claimants with scores below the marginal scores are by design left out from treatment.

II.3 The Data

The nature of Kentucky's WPRS institutions allows the identification of four different samples (which again are depicted in Figure 1): the experimental treated (B) and control (C) samples, which are over a region of random overlap in the distribution of (continuous) profiling scores; and the nonexperimental treated (D) and comparison (A) samples, which are assigned mechanically in and out of treatment following observed discontinuities in the profiling scores.

From June 1994 to October 1996, the period for which we currently have data, 1,236 and 745 claimants are in the experimental treated and control groups, representing 286 PTGs ranging in size from 2 to 54.[3] For the same period, 47,889 and 9,032 claimants form the nonexperimental treated and comparison groups. This means that the experimental design uses only about 2.6 percent of the treated population and 7.6 percent of the untreated population. This relative small experimental sample is a cost of using the randomization at the margin design, which must be weighted against the many virtues described above.

Table 1 presents descriptive statistics for key pre-treatment covariates for each one of the four samples after discarding individuals with problematic information for some covariates of interest.[4] The average continuous profiling scores are 0.83 and 0.79 for the experimental treated and control units, and 0.92 and 0.58 for nonexperimental treatment and comparison units.[5] The

[3] The combination of 87 weeks and 32 local offices give 2,742 potential PTGs. Empty cells, however, for many weeks and local offices gives a final number of 286.

[4] Appendix A describes the sample loss due to listwise deletion of observations present in the data but with problematic values for some key covariates. We discard about 1 percent of the data mainly because of observations with invalid profiling scores.

[5] The corresponding average discrete profiling scores are 15.2, 14.7, 16, and 11.9, respectively.

nonexperimental comparison units present much lower annual earnings ($16,493) than the other groups ($19,000). In terms of schooling and age, all groups show similar mean values for schooling (12 years) and age (37 years), but the nonexperimental treatment and comparison groups differ on most other measures.

II.4. The Regression Discontinuity Groups (RDGs)

Each PTG potentially yields two discontinuities, one at the upper end obtained by using the control observations from the PTG and adding treated observations with higher scores, and one at the lower end obtained by using the treated observations from the PTG and adding untreated observations with lower scores. In the spirit of Rosenbaum (1987) suggestion of using alternative comparison groups to better identify program impacts, we could construct two alternative nonexperimental samples, from above and below, each composed of treated and untreated individuals located in each side of a boundary along the continuous profiling score. Again, as Figure One shows, we match treated and untreated individuals conditional on week and local office within the sample of individuals in groups D and C-A. We call the sample associated with each such discontinuity a regression discontinuity group, or RDGs from above– groups of claimants with at least one treated and one untreated individual in a given office and in a given week, located in each side of the discontinuous point along the continuous profiling score. Similarly, the sample associated with each such discontinuity arising from claimants in groups B-D, and A form the second nonexperimental sample, called RDGs from below. To limit the number of tables in the paper, we present only estimates for the discontinuity from below, but the estimates for the discontinuity from above are available on Smith's web site: http://www-personal.umich.edu/~econjeff/.)

Figure Two shows the asymmetric distribution of PTGs across the 32 local employment offices in the Commonwealth of Kentucky, which reflects the large degree of heterogeneity within the Kentucky economy. The figure reveals a high concentration of PTGs in few local offices. For instance, Northern Kentucky (e.g., Covington, Louisville, and Fern Valley) counts for almost 30 percent of the total number of experimental groups, whereas the western region (e.g., Paducah, Mayfield, and Murray) has less than 3 percent of them, and other areas like Maysville and Danville have none. The average number of PTGs per office is 8.9, ranging in size from 2 to 56. It is clear that those few local offices in which the mass of the experimental observations is located will drive the experimental treatment impacts.

As we wish to evaluate the performance of frequently used RD estimators to the benchmark experimental data, we consider only those RDGs with corresponding PTGs in a given office and in a given week. In addition, we must have additional nonexperimental treatment or comparison observations – the groups A and D from Figure One – to form RDGs. Thus, 272 RDGs from below emerges, ranging in size from 4 to 241. The mean size is 65.5, with a 25th percentile of 36, and a 75th percentile of 83.

II. The RD Approach

Let Y_1 and Y_0 denote the potential outcomes of interest in the treated and untreated states. Let $T_i = 1$ indicate if the individuals are assigned into treatment, and $T_i = 0$ otherwise. The primary object of interest is to estimate the treatment gains $Y_1 - Y_0$ for individuals that receive treatment. Since for any one individual we cannot observe Y_1 and Y_0 simultaneously, the identification of the counterfactual outcome is at the center of the evaluation problem. Instead, we observe $Y = TY_1 + (1-T)Y_0$.

Adopting the potential outcomes framework allows us to illustrate the evaluation problem in the context of a simple linear regression model

$$Y_i = \beta_1 + \beta_2 T_i + u_i \tag{1}$$

where $\{Y_i, T_i\}$ are observed random variables and u_i is the (unobserved) error term. When the binary treatment variable is correlated with the unobservables ($E(u_i \mid T_i) \neq 0$)), the estimated parameter $\hat{\beta}_2$ will not have a causal interpretation.

With sufficient information about the selection process, it is possible to identify causal effects as a direct result of observing treatment assignment rules. For instance, in the simple "sharp" RD design (Trochim 1984), individuals who fall below an observed threshold $\bar{S}$ are mechanically left out from treatment whereas individuals on or above the threshold $\bar{S}$ are assigned into treatment. Thus, this treatment assignment follows a simple deterministic rule $T_i = 1\{S_1 \geq \bar{S}\}$ that can be exploited to identify causal effects, which generates a jump in the probability of assignment at the point $\bar{S}$. If there is a nonzero treatment effect at this point, the assignment rule will induce a discontinuity (jump) in the observed relationship between Y and S at the point at $\bar{S}$. The main idea is to exploit this information for the sample of individuals that are marginally above and below the threshold, $\bar{S} - \varepsilon < S < \bar{S} + \varepsilon$. Because they have essentially the same S , any jump in the outcome Y should be the result of the treatment rather than changes in S because the direct impact of S on the potential outcomes is likely to vary only a little with *S*. Formally, we estimate

$$\begin{aligned} E(Y \mid S = \bar{S} + \varepsilon) - E(Y \mid S = \bar{S} - \varepsilon) = {} & E(Y_1 - Y_0 \mid S = \bar{S} + \varepsilon) \\ & + \underbrace{E(Y_0 \mid S = \bar{S} + \varepsilon) - E(Y_0 \mid S = \bar{S} - \varepsilon)}_{Bias} \end{aligned} \tag{2}$$

where ε is a small number. The first term of the right-hand side of equation (2) is the true treatment effect and the bias term reflects the fact that we have no overlap in our treated and untreated samples and we must use observations that are bit below our cutoff, $\bar{S}$. The appeal of the RD design is that under weak assumptions, this bias term converges to zero. If we assume that $E(Y_0 \mid S)$ is a smooth (continuous) function at $S = \bar{S}$ (Hahn et al, 2001), then the bias term in equation (2) disappears as sample size grows and $\varepsilon \to 0$.[6] Of course, in finite samples this bias term is finite.

III. RD Analysis of the Kentucky WPRS Program

The treatment assignment in the KWPRS program determines simultaneously randomized data (PTGs) along with nonexperimental data that conforms to that of the "sharp" RD assignment mechanism (RDGs). For a given week and a given week, treatment outside the PTGs is assigned based on whether the profiling score *(S)* crosses an observed marginal score $(\bar{S})$. Figure Three graphically represents this unique assignment mechanism for a given RDG. It shows the possible relationship between an outcome of interest (earnings) and the scores that predict the expected benefit receipt duration for two different states of the world. $B(\bar{S})$ represents the relation when UI claimants with the marginal score $(\bar{S})$ are randomly assigned into treatment. $C(\bar{S})$ represents the relation when claimants with the same marginal score are randomly denied access to the reemployment services. $D(S)$ is the relation –only existing for $S > \bar{S}$ – when claimants with high probabilities of benefit exhaustion are mechanically assigned into treatment. Finally, $A(S)$ represents the relationship –only existing for $S < \bar{S}$ –when claimant with low predicted scores of

[6] In many situations, however, the treatment assignment depends on *S* in a stochastic rather than in a deterministic way (e.g., van der Klauuw 2002), resulting in a "fuzzy" RD design.

benefit exhaustion are by design left out from treatment. The positive slopes of $D(S)$ and $A(S)$ reflects the situation that the treatment effects would vary in a deterministic way with *S*.

Whereas the potential gap between $B(\bar{S})$ and $C(\bar{S})$ represents the benchmark experimental treatment impacts, we can also identify nonexperimental treatment impacts by matching the nonexperimental treated sample to the full sample of untreated individuals and comparing the potential gap between $B(\bar{S}) - D(S = \bar{S})$ and $A(S = \bar{S})$ at the B-A discontinuity.

What parameter of interest is representing the gap at the discontinuity point? One can think the marginal (continuous) score as a binary instrument that is only valid at one point. Thus, estimate identifies a local average treatment effect (LATE) (Imbens and Angrist 1994) at the point of discontinuity. Because we have a "sharp" RD design, the LATE parameter is equivalent to the average treatment effect on the treated (ATET) at the discontinuity point. Without some additional assumptions (e.g., common treatment effect), the results do not generalize to identify average treatment effects at other values of *S*.

The RD identification is only possible if the counterfactual earnings vary smoothly with the profiling scores. In the context of the Kentucky WPRS program, it is equivalent to state that nonrandom selection in the RDGs mimics a randomized event in the neighborhood of the marginal profiling scores. Thus, when we use the RD from above (below), we can identify causal effects at the D-C (B-A) discontinuity. As discussed above, comparing that to the experimental treatment impacts constructed using the B and C groups then presumes some amount of smoothness in the estimated mean treatment effect along the profiling score dimension at this point.

A limitation of the RD design is that when treatment impacts vary with the value of the variable that generates the discontinuity, the estimates only applies to the subset of individuals in

the neighborhood around the cut-off point. As we can see later, we take advantage of the unique Kentucky WPRS design that, unlike traditional RD designs, also allows us to estimate heterogeneous treatment effects that vary with *S* over the range of values that have treated and untreated individuals.

IV. The Measurement Framework

IV.1. Parametric Approach

If one knows the true form of the conditional mean function or believes it possible to robustly approximate the true form, one can benefit from additional information contained in observations far from the discontinuity frontier with a parametric framework. A model for individual outcomes is used to describe the causal relationship to be estimated. For the *ith* individual in discontinuity group *j*, we can write

$$\begin{aligned} Y_{ij} &= \delta T_{ij} + g(S_{ij}) + \eta_j + \varepsilon_{ij} \\ T_{ij} &= 1\{S_{ij} \geq \bar{S}_j\} \end{aligned} \tag{3}$$

where Y_{ij} is individual *i*'s outcome, T_{ij} is an indicator variable for treatment status, $g(S_{ij})$ captures the effect of profiling scores on the outcome variable. The term $\eta_j, j = 1,...,J$ denote RDG fixed effects, and ε_{ij} are the error components specific to each individual. We attempted to allow $g(S_{ij})$ to vary by office and/or by week, but failed because of the modest sample sizes of many of our RDGs. In a common effect world, OLS estimation of (3) consistently estimates the common effect. We may also let the outcome variable depend on a set of covariates, which we denote X_{ij}.

The estimated treatment impacts, of course, depend on whether the function $g(\cdot)$ is estimated properly. Theory provides little guidance to choose the correct specification. If this function is correctly specified, this regression-based approach is efficiently using data that are

both close to and far from the discontinuity frontier. On the contrary, when $g(\cdot)$ is not correctly specified the RD estimates will be biased. That is the reason why the empirical literature considers a wide range of alternative specifications of $g(\cdot)$, such as polynomials and splines (e.g., Van der Klaauw 2002).

The sensitiveness of the RD treatment impacts to the specification of $g(\cdot)$ is, however, an unsettle issue in the applied literature. The findings range from no sensitivity (Lemieux et al. 2004, Lalive 2006) to strong sensitivity (Van der Klaauw 2002, DiNardo and Lee 2004). We are better able to illustrate how the bias changes when entering alternative specifications for $g(\cdot)$ because we can evaluate the performance of our $g(\cdot)$ against the experimental results. In addition to implementing simple linear and quadratic functional forms, the order of the polynomial approximation to the population $g(S_{ij})$ function is also selected by the data via penalized cross-validation that minimizes the estimated mean squared error after accounting the degrees of freedom by penalizing models with a large number of coefficients (Akaike, 1970, Blundell and Duncan, 1998).[7] We also minimize the importance of our parametric functional form assumptions by imposing a caliper around the cut-off points on the data we use to estimate the parametric model. Because many approximations (e.g., Taylor series) perform best about a point, the use of a caliper can limit approximation error, and the availability of experimental estimates allow us to address the trade-off between bias and variance for parametric RD models with varying calipers. In addition, we explore the sensitiveness of the regression-based RD estimates to the inclusion of any combination of baseline covariates in the outcome equation. An important policy question that is at the heart of UI profiling programs is whether claimants with high probabilities of benefit exhaustion benefit more from reemployment services than

[7] We do not follow the leave-one-out cross-validation criterion because of some undesirable properties of this approach in the context of parametric selection models (Shao 1993).

those with low probabilities of benefit exhaustion. To address this policy question we take advantage of the unique Kentucky WPRS design that, unlike traditional RD designs, also allows us to estimate heterogeneous treatment effects that vary with *S* over the range of values that have treated and untreated individuals. In the simplest case, we can estimate a version of the outcome equation (3) that interact the scores with the treatment indicator. It results in the estimation of local average treatment effects as a linear function of *S*, and with $\delta(\overline{S})$ replacing δ.

IV.2. Nonparametric Approach

To minimize the potential for misspecification, we turn to nonparametric estimators. A simple estimation strategy is to construct the Wald estimator that identifies a local average treatment effect on the treated at each discontinuity point (Hahn, et al., 2001). To construct the Wald estimate, we calculate the mean differences in outcomes between claimants above and below the discontinuity that defines each RDG and then take the weighted average of these differences, using as weights defined in equation (11).

To implement the Wald estimator we have to decide how wide a caliper use on each side of each discontinuity. As usual, we face a tradeoff between bias and variance in making this choice. A wider window decreases the variance of the estimates by increasing the number of observations used to construct them but (except in special cases) increases the bias. We present three window widths: 0.05 (one profiling score), 0.10 (two profiling scores) or, for comparison purposes, the full sample. Because the distribution of treated and untreated units among the RDGs changes with the window width, different widths result in different observations being used to construct the mean outcomes.

A more flexible approach than the Wald estimator considers assigning different weights to observations in both sides of the discontinuity frontier depending on how close to or far from

the discontinuity frontier are located. This is the data–driven method of Hahn, et al. (2001), which is in one of the most influential papers in the RD literature. The idea is to estimate a kernel-weighted average of the outcome in each side of the discontinuity and then just differentiate the estimates for some fixed bandwidth value of *S*. The simplest kernel approach is referred in the literature as the Nadaraya-Watson (local constant) estimator

$$\tilde{Y}_1^+ = \frac{\sum_i Y_i 1\{S_i \geq \bar{S}\} K_h(S_i - \bar{S})}{\sum_i 1\{S_i \geq \bar{S}\} K_h(S_i - \bar{S})}, \quad \tilde{Y}_0^- = \frac{\sum_i Y_i 1\{S_i < \bar{S}\} K_h(S_i - \bar{S})}{\sum_i 1\{S_i < \bar{S}\} K_h(S_i - \bar{S})} \tag{4}$$

where $K(\cdot)$ is a kernel function, $1(\cdot)$ is an indicator function that equals one if the condition in parenthesis is satisfied and zero otherwise, *h* is the bandwidth parameter, and $\bar{S}$ the cutoff point. These two terms are weighted averages of the dependent variable values for data just to the right and left to the discontinuity point, where the weights decrease to zero with increasing distance to the point of discontinuity at which the kernel is being estimated.[8]

The RD design creates a classic boundary bias problem. The order of the bias of the local constant kernel is *O(h)* at boundary points and $O(h^2)$ at interior points (see Härdle 1990). In our case all the points of estimation are at boundaries, thus the bias problem is exacerbated because of the lack of support in the RD design. To improve over the local constant bias behavior we implement the local linear estimator that accounts for the biased behavior of the conditional expectations at the boundary points (Fan 1992), and shows important bias-reduction properties in the context of the RD approach (Hahn, et al., 2001, Porter 2003).[9] The local linear estimator of observations just to the right of the discontinuity is given by $\hat{\alpha}_0$ where,

[8] Under some conditions, Hahn et al. (2001) demonstrate this procedure is numerically equivalent to an instrumental variable estimator for the regression of Y on T that uses the indicator function as an instrument, applied to the subsample for which $\bar{S} - h \leq S \leq \bar{S} + h$.

[9] In particular, Hahn, et al. (2001) show the asymptotically normality and a rate of convergence equal to $\sqrt{nh_n}$ for the local linear estimator. Porter (2003) shows that polynomial kernel regression not only has the same asymptotic bias

$$\text{argmin}_{\alpha_0,\beta}\sum_{i=1}^{n}1(S_i \geq S)(Y_i - \alpha_0 - \beta(S_i - \bar{S}))^2 K_h(S_i - \bar{S}) \quad (5)$$

where $K(\cdot)$ is the Epanechnikov kernel function which has a bounded support, and *h* is the bandwidth parameter.[10] One additional advantage for local linear regression is, as Lee (2005) notes, that "… (it) can be interpreted as a weighted average treatment effect: those individuals that are more likely to obtain a draw of V near 0 (*S* near $\bar{S}$) receive more weight than those who are unlikely to obtain such a draw" (Lee 2005, pp. 9).

The kernel approach relies on the selection of the optimal bandwidth parameter that achieves the best possible trade-off between bias and variance. We implement the least-square cross-validation approach that is based on the minimization of the "out-of-sample" prediction error (Li and Racine 2004). We implement this approach on each side of the discontinuity separately,

$$CV(h^+) = \frac{1}{n}\sum_{i=1}^{n}1\{S \geq \bar{S}\}\{Y_i - \hat{m}_{i-1,h^+}(S)\}^2; \quad CV(h^-) = \frac{1}{n}\sum_{i=1}^{n}1\{S < \bar{S}\}\{Y_i - \hat{m}_{i-1,h^-}(S)\}^2 \quad (6)$$

where $\hat{m}_{i-1}(\cdot)$ is the smoothed predicted outcome for the observation *ith* when the observation $\{Y_i, S_i\}$ is excluded from the estimation.[11] We check the sensitiveness of the RD estimates to the

---

as typical kernel regression at an interior point of the support, but also in some cases can exhibit further bias reductions. He finds that the local polynomial estimator achieves the optimal convergence rate in the Stone's (1980) sense for estimation of a conditional mean at a point.

[10] The Epanechnikov kernel is given by

$$K_c(x_i, x, h) = \begin{cases} \frac{1}{h}\left(\frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}(\frac{x_i - x}{h})^2\right)\right), if \left|\frac{x_i - x}{h}\right| < \sqrt{5} \\ 0, \text{otherwise} \end{cases},$$

where the range of h is (0, ∞).

[11] If we use the same units to construct the estimator as well as to assess its performance, it will yield a trivial minimum prediction error at $h = 0$. This results, of course, because the best estimate of Y_i is, of course, itself, which occurs in kernel estimation when $h = 0$. The bandwidth search grid equals (0.05, 0.10, …, 2.00) for local linear kernel estimator.

bandwidth parameter by also using bandwidths equal to 0.5 and 1.5 times the value of the cross-validation bandwidth.

Because the rate of convergence for the kernel estimators is much slower than that for the parametric model, the availability of data sets of moderate size is required for the precision of the nonparametric estimates. For this reason, when implementing the kernel-based estimator, we pool the information across all RDGs and re-center the data to a unique discontinuity point by using the "profiling margin of treatment" as the discontinuous variable. It also would allow us to compare this estimator to the parametric one without also changing the pooling decision at the same time.

V. Results

This section measures the extent of the bias for a variety of RD estimators. In most estimations, we present three window widths: 0.05 (one profiling score), 0.10 (two profiling scores) and, for comparison purposes, the full sample. The outcomes of interest are quarterly earnings measured over the 52-week period starting in the first week of the UI claim, and employment in the first quarter after the initial claim.

V.1. The Experimental Estimates

Because the Kentucky program ensures a random assignment only within each PTG and the random assignment ratio differs by PTG, the simple mean differences of outcomes between treated (B) and control units (C) do not estimate the experimental impacts because the rate of assignment to treatment differs across PTG. Thus, following Black et al. (2003a), we estimate two different experimental estimators. First, we run least squares regression of the earnings outcomes on the treatment indicator and a vector of PTG fixed effects to control for differences in expected earnings in the absence of the treatment across PTGs. Thus, we estimate

$$Y_{ij} = \beta T_{ij} + \eta_j + \varepsilon_{ij} \tag{7}$$

where y_{ij} is the outcome for the ith person in the jth PTG, μ_j is a "fixed-effect" for the jth PTG, D_{ij} is the treatment indicator, ε_{ij} is the regression error, and β is the estimate of the impact of treatment, which provides consistent estimates under the standard regression assumption that β is a constant..

One way to think about a PTG is that each PTG is a separate experiment, and the estimate of the treatment effect for each experiment, $\hat{\Delta}_j$, is:

$$\hat{\Delta}_j = \overline{Y}_{1j} - \overline{Y}_{0j} \tag{8}$$

where $\overline{Y}_{1j}$ and $\overline{Y}_{0j}$ are the means for treatment and control groups in the jth PTG. The OLS estimates β as

$$\beta = \sum_j w_j \hat{\Delta}_j \tag{9}$$

where w_j is a weight. In OLS, the weight w_j is given by

$$w_j = \frac{r_j(1-r_j)N_j}{\sum_{k=1}^{286} r_k(1-r_k)N_k}, \tag{10}$$

where N_j is the number of claimants in the jth PTG, and r_j is the probability that a member of jth PTG receives treatment. Thus, OLS weights a PTG more heavily (1) the larger is N_j, the number of claimants, and (2) the closer to 0.5 is the rate of assignment to treatment, r_j. The resulting least squares experimental estimates are $525 for first quarter earnings, $344 for second quarter earnings, $220 for third quarter earnings, $-35 for fourth quarter earnings, and 9.2 percentage points for employment one quarter after treatment.

If the impact of treatment varies across recipients, however, the estimated impact of treatment will vary when the sample varies. Importantly, the number of RDG's differs by the size of the caliper we use. We have 170 and 239 PTGs when we use calipers of width 0.05 and 0.10 from below, which is considerable fewer than the 286 we have for the full experimental estimates. To account for the possibility, we reestimate the experimental estimate to account for heterogeneous treatment effects. The estimator may also be expressed as equation (9) with the weights being given by

$$\tilde{w}_j = \frac{N_{j,T=1}}{\sum_k N_{k,T=1}}, \tag{11}$$

where $N_{k,T=1}$ is the number of treated observations in the *kth* PTG so that the weights are just the proportion of treated units in each PTG. In our experimental sample with 286 PTGs, we cannot reject the hypothesis that the estimated treatment effect is, in fact a constant. When we compare RDD estimators to the experimental estimates, however, we will report bias estimates using both sets of experimental estimates.

V.2 Bias from Parametric Models

Table 2 shows the magnitude of the bias emerging from the basic Equation (3)**.** Each row corresponds to different outcomes of interest. The first column depicts the full experimental impacts as in Black et al. (2003a), and the remaining columns show estimates of the impact of treatment and the bias for three samples: calipers of 0.05, 0.10, and no caliper at all. Within each cell, we present point estimate of the treatment effect, below that we present the standard error of the estimate in parentheses. To the right of the point estimate and its standard error, we present two measures of bias. The first, on the top, uses the fixed-effect experimental estimates with the full sample. The second measure of bias, on the bottom, uses only the experimental

impacts estimated from PTGs with corresponding RDGs using the weights provided in equation (11).

We present estimates from two different models. First, we use OLS to equation (3). Second, we use probit model for our binary employment outcome and a tobit model for our earnings, which accounts for the of earnings at zero. Three main results emerge. First, unlike the parametric models that LaLonde (1986) and Fraker and Maynard (1987) examine, the OLS models replicate the experimental estimates fairly well. Second, for the OLS models, the calipers substantially reduce the bias of the estimators. Both the 0.05 caliper and the 0.10 caliper samples provide estimates with lower bias than using the full sample. Finally, the use of probit for the employment outcome and tobit for the earnings models, particularly, appears to reduce bias.

Table 3 shows the value of including a set of baseline covariates other than *S* in the outcome equation. We consider the same econometric models and same outcomes as in Table 2. For the earnings outcomes the inclusion of the baseline covariates does generally reduce the variance, but generally also increases the bias (although the amount of the variance reduction and bias increase is small). In contrast, for the employment measure the inclusion of the covariates reduces (modestly) the bias. Thus, the inclusion of the baseline covariates appears to have little impact on the estimates.

V.3 Bias from Nonparametric Models

To minimize the potential misspecification of the function $g(S)$ (or $g(S,T)$), we turn to nonparametric estimates. Table 4 shows the bias results for the simple local Wald estimator that compares mean differences in outcomes for observations just above and below the threshold. To construct the estimates we difference the mean outcomes of treated and untreated individuals

within RDG, and the estimates, then, are weighted using the weights given in equation (11). The first column shows the full experimental impacts. The remaining columns show the estimates for three calipers widths: 0.05, 0.10, and the full sample. As before, within each cell there first column provides the estimates and its standard error in parentheses. To the right of the point estimate and its standard error, we present two measures of bias. The first, on the top, uses the fixed-effect experimental estimates with the full sample. The second measure of bias, on the bottom, uses only the experimental impacts estimated from PTGs with corresponding RDGs using the weights provided in equation (11).

Two main patterns emerge. First, as in the regression-based approach, there is evidence that the least bias occurs when the sample set is restricted to those observations within 0.05 or 0.10 of the discontinuity threshold. Second, the bias is somewhat similar to those emerging from the simplest parametric model when expanding the sample sizes, although moving to a non-parametric estimator leaves us with larger estimated standard errors. A small advantage for the regression-based estimates is only observed when looking at the parametric model with baseline covariates (Column 2 of Table 3 versus Columns 2 of Table 4). Overall, the Wald estimates imply substantive inferences similar to those of the parametric estimates; methodologically, we find that, as expected, the performance of the Wald estimator declines more rapidly with the window width.

A second set of nonparametric estimates, reported in Table 5, correspond to Hahn's, et al (2001) one-side local linear kernel regression. We first re-center the data so that all points of discontinuity become zero. We then estimate the necessary bandwidth by using a leave-one-out validation. This method yields the least bias among all RD estimators implemented in this paper. We also observed that the estimator was not too sensitive to the selected bandwidth. Overall, the

kernel-based estimates imply substantive inferences similar to those of the parametric and Wald estimates; methodologically, we find that the performance of the kernel estimator improves over the parametric and Wald estimates.

VI. Evaluating Alternative Discontinuities

The aim of program evaluation is to find the best possible counterfactual for those observations that received treatment. The role of both geography and time dimension on the construction of counterfactuals has been documented in the applied literature. For instance, Friedlander and Robins (1995), Heckman, et al. (1998), and Heckman and Smith (1999) show the quality of the counterfactuals increases when looking at comparison group observations within defined geographical areas. Likewise, the time dimension is relevant to essentially every study that uses panel data methods that relies on variation over time within observations after removing the cross-sectional variation with fixed effects.

The Kentucky WPRS program allows us the opportunity to look into alternative counterfactuals by looking at "neighbors" in two additional dimensions: weeks and local offices. In addition to having neighbors along the profiling score dimension, holding fixed the week and local office, we can also identify neighbors along the local office dimension, holding fixed the profiling score and week, and neighbors along the week dimension, holding fixed the profiling score and the local office. Which of these three counterfactuals best replicates the experimentally determined treatment effects is of substantive and methodological interest.

VI.1. <u>The Geography (local office) Dimension</u>

The geography-based approach seeks to compare treated and untreated individuals living on both sides of a tightly defined geographical area (i.e., similar labor market conditions). As Figure Two illustrates, most of the PTGs are located in a small number of large urban counties that do not

necessarily share an actual geographic border. For this reason, instead of using actual geographic discontinuities, we created a one-dimensional index of similarity among local offices by averaging past annual earnings from high-school white claimants within each employment office. This index has the ability to separate “rich” local labor markets (e.g., Louisville and Fern Valley) from “poor” local labor markets (e.g., Somerset and Murray). The office index ranges from $12,282 to $22,978, with an average of $17,631. It is important to mention that this geographic approach does not come from intrinsic discontinuities in the institutions governing the Kentucky program because it is proxied by a raw index. In this respect, this geographic discontinuity has less priory validity than the score and week discontinuities.

To compare comparable samples, we construct the geographic-based RDGs by matching treated and untreated units filling claims in the same week and with the same (discretized) score in either side of the PTG cell and without imposing any caliper along the index. Thus, for instance, the corresponding geographic-based RDG for the PTG # X (index=15,000) is form by individuals filling claims in week 5, with profiling scores 10, but living in local offices with index above $15,001 for treated individuals and below $15,000 for untreated individuals. As in the case of the week-discontinuity, we repeat this process for each one of the 286 PTGs. From October 1994 to June 1996, the period for which we currently have data, the numbers of valid geographical-based RDGs is 183, which include 3,249 and 1,105 treated and untreated observations.

Table 6 presents the bias for geographic discontinuity. We consider both the simplest OLS model without baseline covariates and the local linear kernel estimator applied to the pooled and centered sample because of their relative better performance in the previous sections. As before, we use two measures of bias: the ones that treat all of the experimental observations

equally (on top) and the ones that restricts the estimate in observations with a RDD (on the bottom).[12]

Some interesting results emerge. In general, we find evidence that geographic discontinuity perform does not perform as well as the score discontinuity in replicating experimental treatment impacts. For instance, we comparing local linear estimator, the score discontinuity outperforms the geographic discontinuity in four out of five outcomes. Nevertheless, relative to the findings of LaLonde (1986) and Fraker and Maynard (1987), the use of the geographic discontinuity provides a reasonable set of estimates.

VI.2 The Time (Weeks) Dimension

A local comparison at the week discontinuity between treated and untreated observations living in the same local offices and with the same probability of benefit exhaustion can be quite informative about the benefits of using the time dimension in the construction of high-quality counterfactuals. Because each PTG consists of individuals with a specific profiling score at a particular location on a particular week, this week-based discontinuity emerges naturally from the Kentucky WPRS institutions, and thus, it is similar to the score discontinuity regarding the strong priors about the validity of the discontinuity.

To construct the week-based RDGs we consider a four-week window on either side of the PTG cell. Within that window, we use all weeks that provided a discontinuity. For instance, the suppose a PTG is composed by claimants from B and C groups living in Elizabethtown, filling claims in week 5, and with (discretized) score 10. The corresponding RDG is formed by claimants from the same local office and with the same (discretized) score, but who were

[12] The weighted experimental estimates emerging from the geography-restricted PTGs are $657, $335, $112, and $-353 for first, second, third, and fourth quarter earnings, and 0.109 for employment in the first quarter after treatment. Likewise, the experimental estimates emerging from the week-restricted PTGs are $607, $265, $130, and $-141 for first, second, third, and fourth quarter earnings, and 0.081for employment in the first quarter after treatment

profiled in weeks (5, 8) for treated individuals, and weeks (1, 4) for untreated individuals. We repeat this process for each one of the 286 PTGs. As a result, 202 RDGs emerge from the full RD data, including 2,242 treated units and 1,564 untreated ones.

We present the results of this exercise in Table 7. While the temporal discontinuity seems to perform well for the employment in the first quarter after the start of benefits in the case of the local linear estimator, the earnings measures are more biased. For the OLS model, each outcome measure is more biased when using the temporal discontinuity than the score discontinuity. Again, relative to the findings of LaLonde (1986) and Fraker and Maynard (1987), the use of the temporal discontinuity credibly replicates the experimental estimates.

VI. Discussion

In this paper, we have investigated the performance of widely used RD estimators using a unique tie-breaking experiment, the Kentucky WPRS program. Our estimates exploit a series of sharp discontinuities in the assignment of the requirement to receive re-employment services inherent in the implementation of the WPRS program. In our context, the nature of the institutions makes this design particularly credible. Our approach follows LaLonde (1986) and Fraker and Maynard (1987) that evaluates the performance of nonexperimental estimators by using as a benchmark experimentally determined treatment impacts.

Taken as a whole, our estimates suggest that RD does in fact credibly replicate the experimental estimates. Both simple parametric and nonparametric models provide estimates that are usually quite similar to the experimental estimates, although our results do indicate that the local linear estimator of Hahn et al. (2001) provides the best estimates. Why do the RD estimates perform so well when the estimators examined by LaLonde (1986) do so poorly?

One obvious answer is that the design itself is much better able to accommodate the complexity of the data than the more traditional econometric estimators. We are able to exploit an important source of information about the probability treatment: the profiling score. This information provides and important source of information for identification.

We also have two other important advantages over the research design that LaLonde (1986) and Fraker and Maynard (1987). First, we know and exploit the geographic location of our recipients. [The work of Smith .... showing the importance of location gets cited here]. Second, our data for both the experimental samples and the RD samples are drawn from the same source: the UI intake forms and administrative information. In contrast, LaLonde had to rely on data from the PSID and the CPS to construct his comparison group [Again, heavily cite Jeff's stuff on the importance of a common instrument.]

Appendix A. Description of Data

The data used for this analysis correspond to the Kentucky WPRS program over the period October 1994 to June 1996. The total number of observations included in the original dataset includes 57,779 UI claimants in the Commonwealth of Kentucky from which 1,981 observations constitutes the experimental sample.

The sample loss due to listwise deletion of observations is about 1 percent (526 observations) of the original dataset: 49 observations correspond to individuals younger than 16 years and older than 90 years; 27 individuals with annual earnings above $100,000; and 450 observations with corrupted profiling scores. None of the deleted observations involves experimental observations.

We keep 835 observations have missing values for the education attainment category. When using covariates, we set their missing education variables to zero, but include a dummy variable indicating that the values are missing. 57,253 UI claimants compose the final data set: 1,981 experimental observations and 55,272 nonexperimental ones.

References

Ashenfelter, O., D. Ashmore and O. Deschenes. "Do Unemployment Insurance Recipients Actively Seek Work? Randomized Trials in Four U.S. States." *Journal of Econometrics*, 125 (2005), 53-75.

Akaike, H. "Statistical Predictor Identification." *Annals of the Institute for Statistical Mathematics*, 22 (1970), 203-217.

Angrist, J. and V. Lavy. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics,* 144 (1999), 533-576.

Black, D., J. Smith, M. Berger, and B. Noel. "Is the Threat of Reemployment Services More Effective than the Services Themselves? Experimental Evidence from the UI System." *American Economic Review,* 93 (2003a): 1313-1327.

Black, D., J. Smith, M. Plesca, and S. Shannon. "Profiling UI Claimants to Allocate Reemployment Services: Evidence and Recommendations for States." Final Report, U.S Department of Labor, 2003b.

Berger, M., D. Black, A. Chandra, and S. Allen. "Kentucky's Statistical Model of Working Profiling for Unemployment Insurance." *Kentucky Journal of Economics and Business*, 16 (1997), 1-18.

Blundell, R. and A. Duncan. "Kernel Regression in Empirical Microeconomics." *Journal of Human Resources*, 33 (1998), 62-87.

Card, D. and P. Levine. "Extended Benefits and the Duration of UI Spells: Evidence from the New Jersey Benefit Program." *Journal of Public Economics,* 78 (2000), 107-38.

Card, D. and D. Lee "Regression Discontinuity Estimation with Random Specification Error." NBER Working Paper T0322, 2006

Chen, S., and W. Van der Klaauw. "The Effect of Disability Insurance on Labor Supply of Older Individuals in the 1990s", unpublished manuscript, 2004.

DiNardo, J., and D. Lee. "Economics Impacts of New Unionization on Private Sector Employers: 1984-2001." *Quarterly Journal of Economics*, 119 (2004), 1383-1442.

Eberts, R, C. O'Leary, and S. Wandner eds. "Targeting Employment Services.", Kalamazoo, MI: W.E Upjohn Institute for Employment Research, 2002.

Fan, J. "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association,* 87 (1992), 998-1004.

Fraker, T. and R. Maynard. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, 22 (1987), 194-227.

Friedlander, D. and P. Robins. “Evaluating Program Evaluations: New Evidence on Commonly used Nonexperimental Methods.” *The American Economic Review*, 85 (1985), 923-937.

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

Härdle, W. Applied Nonparametric Regression, New York, Cambridge University Press, 1990.

Heckman, J. and J. Hotz. “Choosing Among Alternative Nonexperimental Methods for Estimating the Impacts of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association*, 84 (1989), 862-874.

Heckman, J. and J. Smith. “The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies.” *The Economic Journal*, 109 (1999), 313-348.

Heckman, J., H. Ichimura, J. Smith, and P. Todd . “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66 (1998), 1017-1098.

Imbens, G. “Nonparametric Estimation of Average Treatment Effects under Exogeneity.” *The Review of Economics and Statistics*, 86(2004): 4-29

Imbens, G. and J. Angrist. “Identification and Estimation of Local Average Treatment Effects.”, Econometrica, 62 (1994), 467-75.

Kelso, M. “Worker Profiling and Reemployment Services Profiling Methods: Lessons Learned.” UI Occasional Paper 99-5, Washington, D.C: U.S Department of Labor, 1998.

Lalive, R. “How do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach." CESifo Working Paper Series,1765, 2006.

LaLonde, R. “Evaluating the Econometric Evaluation of Training Programs with Experimental Data.” *The American Economics Review*, 76(1986), 604-620

Lee, D. “Randomized Experiments from Nonrandom Selection in U.S House Elections.” *Journal of Econometrics* (Forthcoming). Manuscript, University of California at Berkeley, 2005.

Lemieux, K. Milligan. “Incentive Effects of Social Assistance: A Regression Discontinuity Approach.” NBER Working Paper 10541, 2004.

Li, Q., and J. Racine “Cross-Validated Local Linear Kernel Regression.” *Statistical Sinica*, 14 (2004), 485-512.

Martorell, F. “Do Graduation Exams Matter. A Regression-Discontinuity Analysis of the Impact of Failing the Exit Exam on High School and Post-High School Outcomes.” Manuscript, UC Berkeley, 2004.

Matsudaira, J. “Sinking or Swimming. Evaluation the Impact of English Immersion vs. Bilingual Education on Student Achievement.” Manuscript, University of Michigan, 2004.

McCall, J. "Economics of Information and Job Search." *Quarterly Journal of Economics*, 84 (1970), 113-126.

Meyer, B. “Lessons from the U.S Unemployment Insurance Experiments.” *Journal of Economic Literature*, 33 (1995), 91-131.

Meyer, B. “Unemployment Insurance and Unemployment Spells.” *Econometrica* 58 (1990), 757-782.

Moffitt, R. and W. Nicholson. “The Effect of Unemployment Insurance on Unemployment: The Case of Federal Supplemental Benefits.” *Review of Economics and Statistics*, 64 (1982), 1-11.

Mortensen, D. "Job Search, The Duration of Unemployment, and the Phillips Curve." *American Economic Review*, 60 (1970), 505-517.

O’Leary, C., P. Decker, and S. Wandner. “Reemployment Bonuses and Profiling”, W.E Upjohn Institute Staff Working Paper # 98-51, 1998.

Porter, J. “Estimation on the Regression Discontinuity Model.” Unpublished manuscript, 2003.

Quandt, R.E. “A New Approach to Estimating Switching Regression.” *Journal of the American Statistical Association,* 67 (1972).

Rosenbaum, P. “The Role of a Second Control Group in a Observational Study.” *Statistical Science*, 2 (1987), 292-316.

Roy, A. “Some Thoughts on the Distribution of Earnings.” *Oxford Economic Papers*, 3 (1951), 135-146.

Rubin, D. “Assignment to Treatment Group on the Basis of a Covariate.” *Journal of Education Statistics*, 2 (1977), 1-22.

Shao, J. “Linear Model Selection by Cross-Validation.” Journal of the American Statistical Association, 88 (1993), 486-494.

U.S Department of Labor. “An Analysis of Pooled Evidence from the Pennsylvania and Washington Reemployment Bonus Demonstration.” UI Occasional Paper 92-7, 1992.

U.S Department of Labor. “New Jersey Unemployment Insurance Reemployment Demonstration Project.” UI Occasional Paper 89-3,1989.

Thistlethwaite, D.L. and D.T. Campbell “Regression Discontinuity Analysis: An Alternative to ex post facto Experiment.” *Journal of Educational Psychology*, 51 (1960), 309-17.

Trochim, W. *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Sage Publications, Beverly Hills, 1984.

Van der Klaauw, W. “Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach.” *International Economic Review*, 43 (2002).

Woodbury, S. and R. Spiegelman. “Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois” *American Economic Review*, 77 (1987), 513-530.

Figure 1: Kentucky WRPS Treatment Design

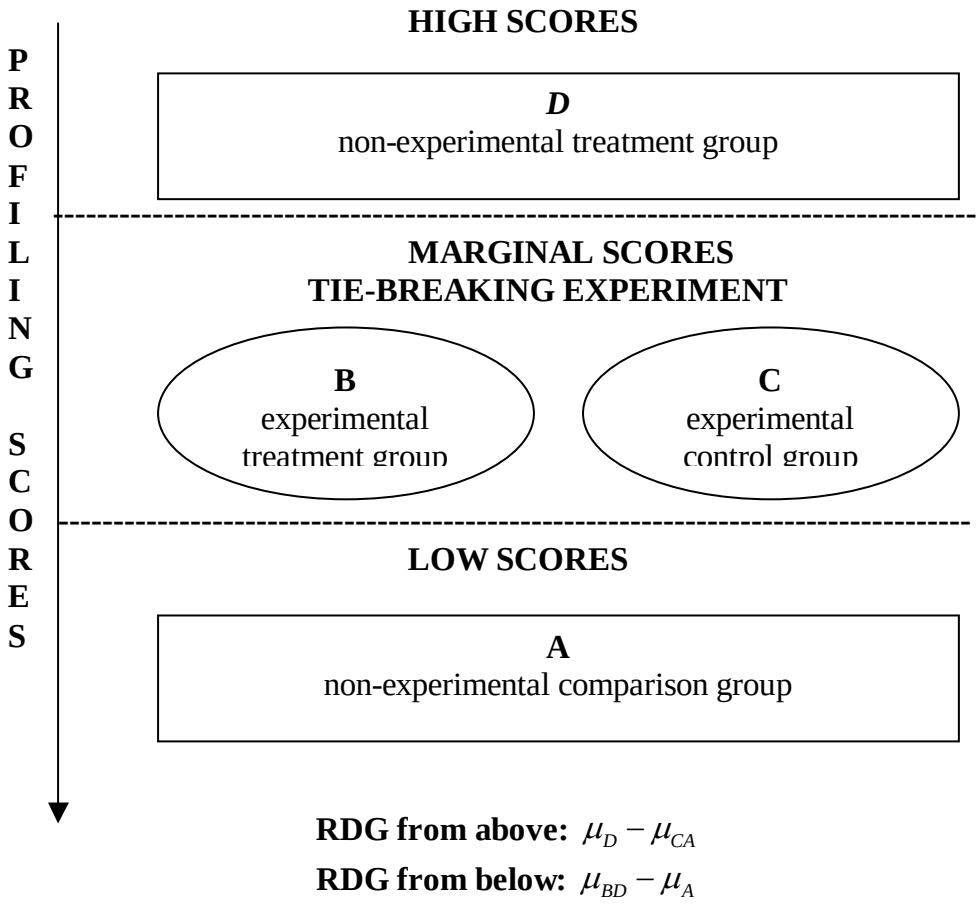

Figure 2: Geographic Distribution of Profiling Tie Groups - PTGs
Kentucky Working and Reemployment Services, October 1994 to June 1996

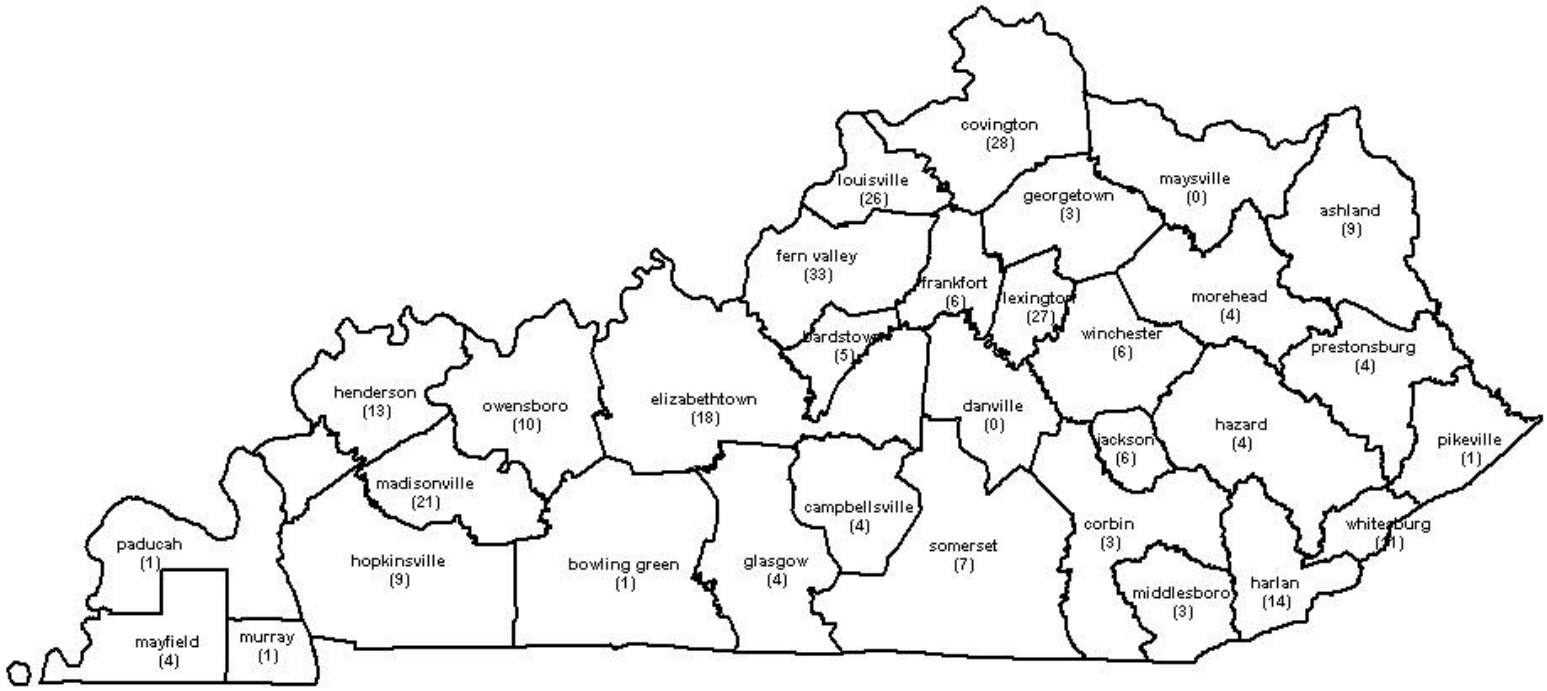

Figure Three: The Kentucky WPRS Regression Discontinuity Design

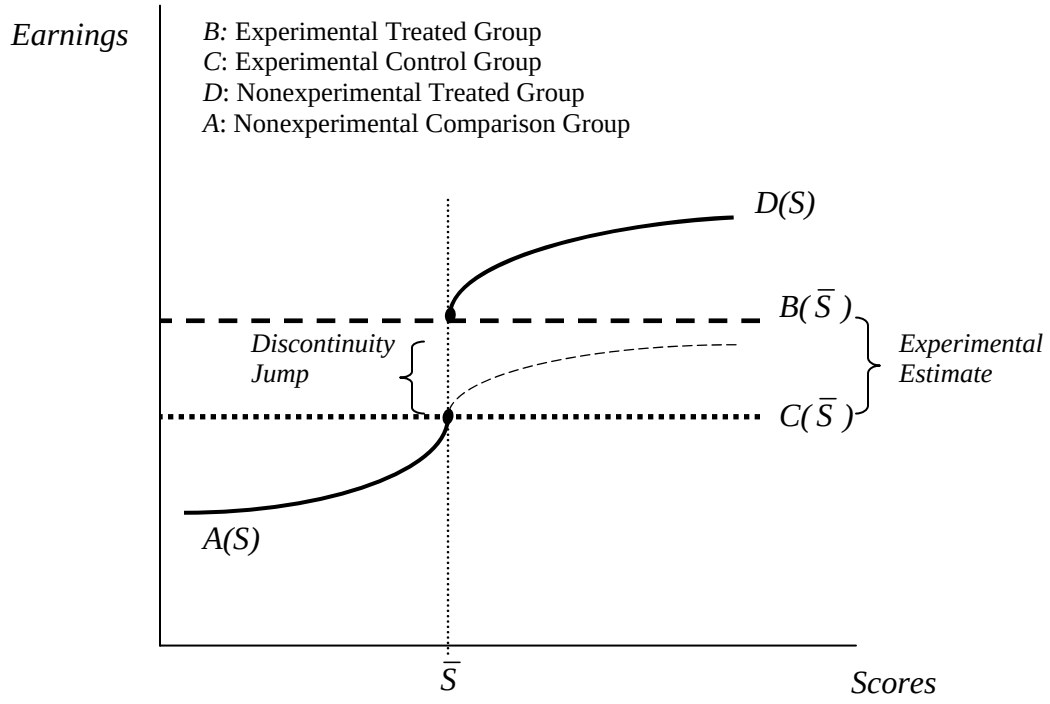

Table 1: Summary Statistics
Kentucky Working and Reemployment Services, October 1994 to June 1996

	Experimental Sample Set			Nonexperimental Sample Set		
	Treated *(B)*	Control *(C)*	p-values for test of differences in means	Treated *(D)*	Comparison *(A)*	p-values for test of differences in means
Profiling score	0.83	0.79	0.89	0.92	0.58	0.00
	(0.22)	(0.21)		(0.27)	(0.23)	
Annual earnings	$19046	$19758	0.66	$18942	$16493	0.00
	(13636)	(13676)		(13343)	(12677)	
1st quarter earnings	$4555	$5008	0.82	$4515	$3975	0.00
	(3815)	(4072)		(3717)	(3636)	
2nd quarter earnings	$4461	$4680	0.84	$4624	$3853	0.00
	(3832)	(3745)		(3805)	(3566)	
3rd quarter earnings	$4988	$4967	0.81	$4821	$4169	0.00
	(3789)	(3514)		(3562)	(3503)	
4th quarter earnings	$5131	$5102	0.39	$4980	$4494	0.00
	(3731)	(3608)		(3606)	(3383)	
Years of schooling	12.5	12.3	0.22	12.3	12.4	0.96
	(2.1)	(2.0)		(2.0)	(1.9)	
1st quarter employment (%)	92	91	0.49	92	89	0.00
	(0.26)	(0.27)		(0.26)	(0.31)	
Less than high school (%)	15	18	0.49	16	13	0.00
	(0.35)	(0.38)		(0.36)	(0.33)	
Bachelor degree (%)	5.3	5.2	0.30	4.5	4.1	0.01
	(0.22)	(0.2)		(0.2)	(0.2)	
Graduate studies (%)	2.8	2.1	0.97	2.5	1.8	0.00
	(0.16)	(0.14)		(0.15)	(0.13)	
Age	37.0	37.0	0.71	37.3	36.6	0.00
	(11)	(10.8)		(11.2)	(11.4)	
Percent females	42.9	39.6	0.04	44.2	41	0.00
	(0.4)	(0.4)		(0.5)	(0.5)	
Percent whites	88.9	91.6	0.78	91.5	90.3	0.08
	(0.3)	(0.2)		(0.3)	(0.3)	
Percent blacks	10.5	8.1	0.90	8.1	9.2	0.13
	(0.3)	(0.3)		(0.3)	(0.3)	
N	1,236	745		46,270	9,002	

Notes: Standard deviations are given in parenthesis. Means are unweighted. Test for differences in means for the experimental sample (B versus C) are based on a linear regression that conditions on a treatment dummy variable and on the PTGs. Test for differences in means for the nonexperimental sample (D versus A) are based on a linear regression that conditions on a treatment dummy variable and on local office and week.

Table 2: Bias of Parametric Regression Estimates without Baseline Covariates, Kentucky WPRS, October 1994 to June 1996

	Experimental Estimates	0.05 caliper Estimate	0.05 caliper Bias	0.10 caliper Estimate	0.10 caliper Bias	No caliper Estimate	No caliper Bias
OLS estimates							
1st quarter employment	9.2	7.1	-2.1	7.6	-1.6	2.2	-7.0
	(2.6)	(2.8)	-2.4	(1.9)	-3.0	(1.3)	-8.4
1st quarter earnings	525	377	-148	483	-42	221	-304
	(192)	(209)	-334	(139)	-255	(83)	-486
2nd quarter earnings	344	207	-137	247	-97	104	-240
	(161)	(169)	-206	(133)	-190	(84)	-321
3rd quarter earnings	220	96	-124	180	-40	-20	-240
	(181)	(191)	-241	(130)	-174	(86)	-328
4th quarter earnings	-36	-84	-48	103	139	-25	11
	(176)	(184)	-60	(127)	119	(86)	33
Probit/Tobit models (marginal effects reported)							
1st quarter employment	9.2	7.1	-2.1	7.6	-1.6	2.2	-7.0
	(2.6)	(2.8)	-2.4	(1.9)	-3.0	(1.3)	-8.4
1st quarter earnings	525	954	429	1,120	595	421	-104
	(192)	(395)	243	(267)	382	(168)	-286
2nd quarter earnings	344	431	87	540	196	197	-147
	(161)	(297)	18	(223)	103	(145)	-228
3rd quarter earnings	220	232	12	395	175	30	-190
	(181)	(312)	-105	(212)	41	(143)	-278
4th quarter earnings	-36	-79	-43	259	295	-3	33
	(176)	(308)	-55	(212)	275	(147)	55
PTG/RDG	286	170		239		272	
N	1,981	1,337		2,888		11,826	

Notes: Non-experimental impact estimates are from the model

$$Y_{ij} = \delta T_{ij} + g(S_{ij}) + \varepsilon_{ij},$$

where T is the treatment indicator and the order of the polynomial approximation to g(S) is selected by Akaike's penalized functions. Standard errors are in parenthesis. Estimated bias is equal to the difference between non-experimental and experimental impacts. The first bias is based on the full experimental estimates, whereas the second one uses only PTGs with corresponding valid RDGs with weights given in equation (11).

Table 3: Bias of Parametric Regression Estimates with Baseline Covariates,, Kentucky WPRS, October 1994 to June 1996

	Experimental Estimates	0.05 caliper Estimate	0.05 caliper Bias	0.10 caliper Estimate	0.10 caliper Bias	No caliper Estimate	No caliper Bias
OLS estimates							
1st quarter employment	9.2	7.2	-2.0	7.8	-1.4	3.2	-6.0
	(2.6)	(2.8)	-2.3	(1.9)	-2.8	(1.3)	-7.4
1st quarter earnings	525	367	-158	436	-89	115	-410
	(192)	(203)	-344	(133)	-302	(84)	-592
2nd quarter earnings	344	162	-182	185	-159	-17	-361
	(161)	(163)	-251	(127)	-252	(82)	-442
3rd quarter earnings	220	35	-185	12 4	-96	-167	-387
	(181)	(183)	-302	(124)	-230	(83)	-475
4th quarter earnings	-36	-124	-88	34	70	-101	-65
	(176)	(173)	-100	(120)	50	(82)	-43
Probit/Tobit models							
1st quarter employment	9.2	7.4	-1.8	7.9	-1.3	3.2	-6.0
	(2.6)	(2.8)	-2.1	(1.9)	-2.7	(1.3)	-7.4
1st quarter earnings	525	976	451	1054	529	318	-207
	(192)	(383)	265	(254)	316	(161)	-389
2nd quarter earnings	344	369	25	476	132	84	-260
	(161)	(284)	-44	(213)	39	(141)	-341
3rd quarter earnings	220	184	-36	332	112	-126	-346
	(181)	(296)	-153	(201)	-22	(139)	-434
4th quarter earnings	-36	-136	-100	158	194	-165	-129
	(176)	(289)	-112	(199)	174	(141)	-107
PTG/RDG	286	170		239		272	
N	1,981	1,337		2,888		11,826	

Notes: Non-experimental impact estimates are from the model

$$Y_{ij} = X_{ij}\beta + \delta T_{ij} + g(S_{ij}) + \varepsilon_{ij},$$

where T is the treatment indicator and the order of the polynomial approximation to g(S) is selected by Akaike's penalized functions. X_{ij} includes age and age squared, education, four quarter of earnings before unemployment, dummy variables indicating the worker is female, white, or black, the interaction fo the profiling score and the dummy variable for the worker being white, and interaction of the profiling score and an indicator that the worker female. Standard errors are in parenthesis. Estimated bias is equal to the difference between non-experimental and experimental impacts. The first bias is based on the full experimental estimates, whereas the second one uses only PTGs with corresponding valid RDGs with weights given in equation (11).

Table 4: Bias of Wald Estimates, Kentucky WPRS, October 1994 to June 1996

		0.05 caliper		0.10 caliper		No caliper	
	Experimental Estimates	**Estimate**	**Bias**	**Estimate**	**Bias**	**Estimate**	**Bias**
1st quarter employment	9.2	6.7	-2.5	6.6	-2.6	1.4	-7.8
	(2.6)	(3.9)	-2.8	(2.6)	-4.0	(1.4)	-9.2
1st quarter earnings	525	423	-102	530	5	298	-227
	(192)	(291)	-288	(191)	-208	(93)	-440
2nd quarter earnings	344	208	-136	243	-101	127	-217
	(161)	(236)	-205	(181)	-194	(98)	-310
3rd quarter earnings	220	120	-100	242	22	107	-113
	(181)	(267)	-217	(179)	-112	(97)	-247
4th quarter earnings	-36	-64	-28	166	202	172	208
	(176)	(256)	-40	(175)	182	(94)	188
PTG/RDG	286	170		239		272	
N	1,981	1,337		2,888		11,826	

Notes: The Wald estimates are estimated by mean differences applied separately to each RDG and then weighting up the estimates using as weights given in equation (11). Standard errors are given in parentheses. Estimated bias is equal to the difference between non-experimental and experimental impacts. The first bias is based on the full experimental estimates, whereas the second one uses only PTGs with corresponding valid RDGs with weights given in equation (11).

Table 5: Bias of Local Linear Estimator, Kentucky WPRS, October 1994 to June 1996

		$0.5*h^{OPT}$		h^{OPT}		$1.5*h^{OPT}$	
	Experimental Estimates	**Estimate**	**Bias**	**Estimate**	**Bias**	**Estimate**	**Bias**
1st quarter	9.2	8.7	-0.5	7.1	-2.1	7.1	-2.9
employment	(2.6)	(2.2)	-1.9	(1.7)	-3.5	(1.5)	-4.3
1st quarter	525	496	-29	491	-34	452	-73
earnings	(192)	(128)	-211	(104)	-216	(100)	-255
2nd quarter	344	308	-36	337	-7	295	-49
earnings	(161)	(127)	-117	(99)	-88	(102)	-130
3rd quarter	220	306	86	235	15	223	3
earnings	(181)	(114)	-2	(104)	-73	(100)	-85
4th quarter	-36	250	286	188	224	209	245
earnings	(176)	(140)	308	(114)	246	(110)	267
N	1,981	11,824		11,824		11,824	

Notes: Local linear regression uses the Epanechnikov kernel function with cross-validated bandwidths. Bootstrapped standard errors are given in parentheses. It is based on 500 replications. Estimated bias is equal to the difference between non-experimental and experimental impacts. The first bias is based on the full experimental estimates, whereas the second one uses only PTGs with corresponding valid RDGs with weights given in equation (11).

Table 6: Geographic Discontinuities, Kentucky WPRS, October 1994 to June 1996

		OLS model		Local Linear	
	Experimental Estimates	**Estimate**	**Bias**	**Estimate**	**Bias**
1st quarter	9.2	3.5	-5.7	4.5	-4.7
employment	(2.6)	(1.8)	-7.4	(2.3)	-6.4
1st quarter	525	325	-200	270	-255
earnings	(192)	(110)	-332	(135)	-387
2nd quarter	344	296	-48	210	-134
earnings	(161)	(118)	-39	(143)	-125
3rd quarter	220	164	-56	-11	-231
earnings	(181)	(122)	52	(150)	-123
4th quarter	-36	-20	16	-246	-210
earnings	(176)	(125)	333	(162)	107
N	1,981	4,354		4,354	

Notes: Non-experimental impact estimates are from the model

$$Y_{ij} = \delta T_{ij} + g(S_{ij}) + \varepsilon_{ij},$$

where T is the treatment indicator and the order of the polynomial approximation to g(S) is selected by Akaike's penalized functions. Standard errors are in parenthesis. Estimated bias is equal to the difference between non-experimental and experimental impacts. : Local linear regression uses the Epanechnikov kernel function with cross-validated bandwidths. Bootstrapped standard errors are given in parentheses. It is based on 500 replications. Estimated bias is equal to the difference between non-experimental and experimental impacts. The first bias is based on the full experimental estimates, whereas the second one uses only PTGs with corresponding valid RDGs with weights given in equation (11).

Table 7: Temporal Discontinuities, Kentucky WPRS, October 1994 to June 1996

		OLS estimates		Local Linear, 3 week bandwidth		Local Linear, 4 week bandwidth	
	Experimental Estimates	**Estimate**	**Bias**	**Estimate**	**Bias**	**Estimate**	**Bias**
1st quarter	9.2	2.5	-6.7	11.3	2.1	7.2	-2.0
employment	(2.6)	(1.6)	-5.6	(5.9)	3.2	(4.4)	-0.9
1st quarter	525	144	-381	1044	519	859	334
earnings	(192)	(96)	-463	(419)	437	(290)	252
2nd quarter	344	-139	-483	419	75	323	-21
earnings	(161)	(108)	-404	(397)	154	(253)	58
3rd quarter	220	-166	-386	425	205	278	58
earnings	(181)	(113)	-296	(430)	295	(290)	148
4th quarter	-36	-168	-132	248	284	256	292
earnings	(176)	(113)	-27	(431)	389	(298)	397
N	1,981	3,806		3,806		3,806	

Notes: Non-experimental impact estimates are from the model

$$Y_{ij} = \delta T_{ij} + g(S_{ij}) + \varepsilon_{ij},$$

where T is the treatment indicator and the order of the polynomial approximation to g(S) is selected by Akaike's penalized functions. Standard errors are in parenthesis. Estimated bias is equal to the difference between non-experimental and experimental impacts. : Local linear regression uses the Epanechnikov kernel function with cross-validated bandwidths. Bootstrapped standard errors are given in parentheses. It is based on 500 replications. Estimated bias is equal to the difference between non-experimental and experimental impacts. The first bias is based on the full experimental estimates, whereas the second one uses only PTGs with corresponding valid RDGs with weights given in equation (11).